

*Babatunde A. Ogunnaike*

---

# **Random Phenomena**

*Fundamentals and Engineering Applications of  
Probability & Statistics*

---

## Random Phenomena

Fundamentals and Engineering Applications of Probability & Statistics

---

*I frame no hypothesis; for whatever is not deduced from the phenomenon is to be called a hypothesis; and hypotheses, whether metaphysical or physical, whether of occult qualities or mechanical, have no place in experimental philosophy.*

Sir Isaac Newton (1642–1727)

---

## ***In Memoriam***

In acknowledgement of the debt of birth I can never repay, I humbly dedicate this book to the memory of my father, my mother, and my statistics mentor at Wisconsin.

### **Adesijibomi Ogundero Ogunnaike** 1922–2002

*Some men fly as eagles free  
But few with grace to the same degree  
as when you rise upward to fly  
to avoid sparrows in a crowded sky*

### **Ayoola Oluronke Ogunnaike** 1931–2005

*Some who only search for silver and gold  
Soon find what they cannot hold;  
You searched after God's own heart,  
and left behind, too soon, your pilgrim's chart*

### **William Gordon Hunter** 1937–1986

*See what ripples great teachers make  
With but one inspiring finger  
Touching, once, the young mind's lake*



---

## Preface

In an age characterized by the democratization of quantification, where data about every conceivable phenomenon is available somewhere and easily accessible from just about anywhere, it is becoming just as important that the “educated” person *also* be conversant with how to handle data, and be able to understand what the data say as well as what they don’t say. Of course, this has always been true of scientists and engineers—individuals whose profession requires them to be involved in the acquisition, analysis, interpretation and exploitation of data in one form or another; but it is even more so now. Engineers now work in non-traditional areas ranging from molecular biology to finance; physicists work with material scientists and economists; and the problems to be solved continue to widen in scope, becoming more interdisciplinary as the traditional disciplinary boundaries disappear altogether or are being restructured.

In writing this book, I have been particularly cognizant of these basic facts of 21<sup>st</sup> century science and engineering. And yet while most scientists and engineers are well-trained in problem formulation and problem solving when *all* the entities involved are considered deterministic in character, many remain uncomfortable with problems involving random variations, if such problems cannot be idealized and reduced to the more familiar “deterministic” types. Even after going through the usual one-semester course in “Engineering Statistics,” the discomfort persists. Of all the reasons for this circumstance, the most compelling is this: most of these students tend to perceive their training in statistics more as a set of instructions on *what* to do and *how* to do it, than as a training in fundamental principles of random phenomena. Such students are then uncomfortable when they encounter problems that are not quite similar to those covered in class; they lack the fundamentals to attack new and unfamiliar problems. The purpose of this book is to address this issue directly by presenting basic fundamental principles, methods, and tools for formulating and solving engineering problems that involve randomly varying phenomena. The premise is that by emphasizing fundamentals and basic principles, and then illustrating these with examples, the reader will be better equipped to deal with a range of problems wider than that explicitly covered in the book. This important point is expanded further in Chapter 0.

## Scope and Organization

Developing a textbook that will achieve the objective stated above poses the usual challenge of balancing breadth and depth—an optimization problem with no unique solution. But there is also the additional constraint that the curriculum in most programs can usually only accommodate a one-semester course in engineering statistics—if they can find space for it at all. As all teachers of this material know, finding a universally acceptable compromise solution is impossible. What this text offers is enough material for a two-semester introductory sequence in probability and statistics for scientists and engineers, and with it, the flexibility of several options for using the material. We envisage the following three categories, for which more detailed recommendations for coverage will be provided shortly:

- Category I: The two-semester undergraduate sequence;
- Category II: The traditional one-semester undergraduate course;
- Category III: The one-semester beginning graduate course.

The material has been tested and refined over more than a decade, in the classroom (at the University of Delaware; at the African University of Science and Technology (AUST), in Abuja, Nigeria; at the African Institute of Mathematical Sciences (AIMS) in Muizenberg, South Africa), and in short courses presented to industrial participants at DuPont, W. L. Gore, SIEMENS, the Food and Drugs Administration (FDA), and many others through the University of Delaware's Engineering Outreach program.

The book is organized into 5 parts, after a brief prelude in Chapter 0 where the book's organizing principles are expounded. Part I (Chapters 1 and 2) provides foundational material for understanding the fundamental nature of random variability. Part II (Chapters 3–7) focuses on probability. Chapter 3 introduces the fundamentals of probability theory, and Chapters 4 and 5 extend these to the concept of the random variable and its distribution, for the single and the multidimensional random variable. Chapter 6 is devoted to random variable transformations, and Chapter 7 contains the first of a trilogy of case studies, this one devoted to two problems with substantial historical significance.

Part III (Chapters 8–11) is devoted entirely to developing and analyzing probability models for specific random variables. The distinguishing characteristics of the presentation in Chapters 8 and 9, respectively for discrete and continuous random variables, is that each model is developed from underlying phenomenological mechanisms. Chapter 10 introduces the idea of information and entropy as an alternative means of determining appropriate probability models when only partial knowledge is available about the random variable in question. Chapter 11 presents the second case study, on in-vitro fertilization (IVF), as an application of probability models. The chapter illustrates the development, validation, and use of probability modeling on a contemporary problem with significant practical implications.

The core of statistics is presented in Part IV (Chapters 12–20). Chapter 12 lays the foundation with an introduction to the concepts and ideas behind statistics, before the coverage begins in earnest in Chapter 13 with sampling theory, continuing with statistical inference, estimation and hypothesis testing, in Chapters 14 and 15, and regression analysis in Chapter 16. Chapter 17 introduces the important but oft-neglected issue of probability model validation, while Chapter 18 on nonparametric methods extends the ideas of Chapters 14 and 15 to those cases where the usual probability model assumptions (mostly the normality assumption) are invalid. Chapter 19 presents an overview treatment of design of experiments. The third and final set of case studies is presented in Chapter 20 to illustrate the application of various aspects of statistics to real-life problems.

Part V (Chapters 21–23) showcases the “application” of probability and statistics with a hand-selected set of “special topics:” reliability and life testing in Chapter 21, quality assurance and control in Chapter 22, and multivariate analysis in Chapter 23. Each has roots in probability and statistics, but all have evolved into *bona fide* subject matters in their own rights.

## Key Features

Before presenting suggestions of how to cover the material for various audiences, I think it is important to point out some of the key features of the textbook.

**1. Approach.** This book takes a more fundamental, “first-principles” approach to the issue of dealing with random variability and uncertainty in engineering problems. As a result, for example, the treatment of probability distributions for random variables (Chapters 8–10) is based on a derivation of each model from phenomenological mechanisms, allowing the reader to see the subterranean roots from which these probability models sprang. The reader is then able to see, for instance, how the Poisson model arises either as a limiting case of the binomial random variable, or from the phenomenon of observing in finite-sized intervals of time or space, rare events with low probabilities of occurrence; or how the Gaussian model arises from an accumulation of small random perturbations.

**2. Examples and Case Studies.** This fundamental approach note above is integrated with practical applications in the form of a generous amount of examples but also with the inclusion of three chapter-length application case studies, one each for probability, probability distributions, and statistics. In addition to the usual traditional staples, many of the in-chapter examples have been drawn from non-traditional applications in molecular biology (e.g., DNA replication origin distributions; gene expression data, etc.), from finance and business, and from population demographics.

**3. Computers, Computer Software, On-line resources.** As expanded further in the Appendix, the availability of computers has transformed the teaching and learning of probability and statistics. Statistical software packages are now so widely available that many of what used to be staples of traditional probability and statistics textbooks—tricks for carrying out various computations, approximation techniques, and especially printed statistical tables—are now essentially obsolete. All the examples in this book were carried out with MINITAB, and I fully expect each student and instructor to have access to one such statistical package. In this book, therefore, we depart from tradition and do *not* include any statistical tables. Instead, we have included in the Appendix a compilation of useful information about some popular software packages, on-line electronic versions of statistical tables, and a few other on-line resources such as on-line electronic statistics handbooks, and websites with data sets.

**4. Questions, Exercises, Application Problems, Projects.** No one feels truly confident about a subject matter without having tackled (and solved!) some problems; and a useful textbook ought to provide a good selection that offers a broad range of challenges. Here is what is available in this book:

- *Review Questions:* Found at the end of each chapter (with the exception of the chapters on case studies), these are short, specific questions designed to test the reader's basic comprehension. If you can answer all the review questions at the end of each chapter, you know and understand the material; if not, revisit the relevant portion and rectify the revealed deficiency.
- *Exercises:* are designed to provide the opportunity to master the mechanics behind a single concept. Some may therefore be purely "mechanical" in the sense of requiring basic computations; some may require filling in the steps deliberately "left as an exercise to the reader;" some may have the flavor of an application; but the focus is usually a single aspect of a topic covered in the text, or a straightforward extension thereof.
- *Application Problems:* are more substantial practical problems whose solutions usually require integrating various concepts (some obvious, some not) and deploying the appropriate set of tools. Many of these are drawn from the literature and involve real applications and actual data sets. In such cases, the references are provided, and the reader may wish to consult some of them for additional background and perspective, if necessary.
- *Project assignments:* allow deeper exploration of a few selected issues covered in a chapter, mostly as a way of extending the coverage and also to provide opportunities for creativity. By definition, these involve a significant amount of work and also require report-writing. This book offers a total of nine such projects. They are a good way for students to

learn how to plan, design, and execute projects and to develop writing and reporting skills. (Each graduate student that has taken the CHEG 604 and CHEG 867 courses at the University of Delaware has had to do a term project of this type.)

**5. Data Sets.** All the data sets used in each chapter, whether in the chapter itself, in an example, or in the exercises or application problems, are made available on-line and on CD.

### Suggested Coverage

Of the three categories mentioned earlier, a methodical coverage of the entire textbook is only possible for Category I, in a two-semester undergraduate sequence. For this group, the following is one possible approach to dividing the material up into instruction modules for each semester:

- **First Semester**

*Module 1 (Foundations):* Chapters 0–2.

*Module 2 (Probability):* Chapters 3, 4, 5 and 7.

*Module 3 (Probability Models):* Chapter 8<sup>1</sup> (omit detailed derivations and Section 8.7.2), Chapter 9<sup>1</sup> (omit detailed derivations), and Chapter 11<sup>1</sup> (cover Sections 11.4 and 11.5 selectively; omit Section 11.6).

*Module 4 (Introduction to Statistics/Sampling):* Chapters 12 and 13.

*Module 5 (Statistical Inference):* Chapter 14<sup>1</sup> (omit Section 14.6), Chapter 15<sup>1</sup> (omit Sections 15.8 and 15.9), Chapter 16<sup>1</sup> (omit Sections 16.4.3, 16.4.4, and 16.5.2), and Chapter 17.

*Module 6 (Design of Experiments):* Chapter 19<sup>1</sup> (cover Sections 19.3–19.4 lightly; omit Section 19.10) and Chapter 20.

- **Second Semester**

*Module 7 (Probability and Models):* Chapters 6 (with *ad hoc* reference to Chapters 4 and 5); Chapters 8<sup>2</sup> and 9<sup>2</sup> (include details omitted in the first semester), Chapter 10.

*Module 8 (Statistical Inference):* Chapter 14<sup>2</sup> (Bayesian estimation, Section 14.6), Chapter 15<sup>2</sup> (Sections 15.8 and 15.9), Chapter 16<sup>2</sup> (Sections 16.4.3, 16.4.4, and 16.5.2), and Chapter 18.

*Module 9 (Applications):* Select one of Chapter 21, 22 or 23. (For chemical engineers, and anyone planning to work in the manufacturing industry, I recommend Chapter 22.)

With this as a basic template, other variations can be designed as appropriate.

For example, those who can only afford one semester (Category II) may adopt the first semester suggestion above, to which *I recommend adding Chapter 22 at the end.*

The beginning graduate one-semester course (Category III) may also be based on the first semester suggestion above, but with the following additional recommendations: (i) cover of all the recommended chapters *fully*; (ii) add Chapter 23 on multivariate analysis; and (iii) in lieu of a final examination, assign at least one, possibly two, of the nine projects.

This will make for a hectic semester, but graduate students should be able to handle the workload.

A second, perhaps more straightforward, recommendation for a two-semester sequence is to devote the first semester to Probability (Chapters 0–11), and the second to Statistics (Chapters 12–20) along with one of the three application chapters.

### Acknowledgments

Pulling off a project of this magnitude requires the support and generous assistance of many colleagues, students, and family. Their genuine words of encouragement and the occasional (innocent and not-so-innocent) inquiry about the status of “the book” all contributed to making sure that this potentially endless project was actually finished. At the risk of leaving someone out, I feel some deserve particular mention. I begin with, in alphabetical order, Marc Birtwistle, Ketan Detroja, Claudio Gelmi (Chile), Mary McDonald, Vinay Prasad (Alberta, Canada), Paul Taylor (AIMS, Muizenberg, South Africa), and Carissa Young. These are colleagues, former and current students, and postdocs, who patiently waded through many versions of various chapters, offered invaluable comments and caught many of the manuscript errors, typographical and otherwise. It is a safe bet that the manuscript still contains a random number of these errors (few and Poisson distributed, I hope!) but whatever errors remain are my responsibility. I encourage readers to let me know of the ones they find.

I wish to thank my University of Delaware colleagues, Antony Beris and especially Dion Vlachos, with whom I often shared the responsibility of teaching CHEG 867 to beginning graduate students. Their insight into what the statistics component of the course should contain was invaluable (as were the occasional Greek lessons!). Of my other colleagues, I want to thank Dennis Williams of Basel, for his interest and comments, and then single out former fellow “DuPonters” Mike Piovoso, whose fingerprint is recognizable on the illustrative example of Chapter 23, Rafi Sela, now a Six-Sigma Master Black Belt, Mike Deaton of James Madison University, and Ron Pearson, whose near-encyclopedic knowledge never ceases to amaze me. Many of the ideas, problems and approaches evident in this book arose from those discussions and collaborations from many years ago. Of my other academic colleagues, I wish to thank Carl Laird of Texas A & M for reading some of the chapters, Joe Qin of USC for various suggestions, and Jim Rawlings of Wisconsin with whom I have carried on a long-running discussion about probability and estimation because of his own interests and expertise in this area. David Bacon

and John MacGregor, pioneers in the application of statistics and probability in chemical engineering, deserve my thanks for their early encouragement about the project and for providing the occasional commentary. I also wish to take this opportunity to acknowledge the influence and encouragement of my chemical engineering mentor, Harmon Ray. I learned more from Harmon than he probably knew he was teaching me. Much of what is in this book carries an echo of his voice and reflects the Wisconsin tradition.

I must not forget my gracious hosts at the École Polytechnique Fédérale de Lausanne (EPFL), Professor Dominique Bonvin (*Merci pour tout, mon ami*) and Professor Vassily Hatzimanikatis (*Ευχαριστώ πολὺ παλιοφίλε:* “Efharisto poli paliofile”). Without their generous hospitality during the months from February through July 2009, it is very likely that this project would have dragged on for far longer. I am also grateful to Michael Amrhein of the Laboratoire d’Automatique at EPFL, and his graduate student, Paman Gujral, who both took time to review several chapters and provided additional useful references for Chapter 23. My thanks go to Allison Shatkin and Marsha Pronin of CRC Press/Taylor and Francis for their professionalism in guiding the project through the various phases of the editorial process all the way to production.

And now to family. Many thanks are due to my sons, Damini and Deji, who have had cause to use statistics at various stages of their (still on-going) education: each read and commented on a selected set of chapters. My youngest son, Makinde, still too young to be a proofreader, was nevertheless solicitous of my progress, especially towards the end. More importantly, however, just by “showing up” when he did, and *how*, he confirmed to me without meaning to, that he is a natural-born Bayesian. Finally, the debt of thanks I owe to my wife, Anna, is difficult to express in a few words of prose. She proofread many of the chapter exercises and problems with an incredible eye, and a sensitive ear for the language. But more than that, she knows well what it means to be a “book widow”; without her forbearance, encouragement, and patience, this project would never have been completed.

Babatunde A. Ogunnaike  
Newark, Delaware  
Lausanne, Switzerland

April 2009

x

---

## ***List of Figures***

|     |   |     |
|-----|---|-----|
| 1.1 | Histogram for $Y_A$ data . . . . .  | 19  |
| 1.2 | Histogram for $Y_B$ data . . . . .  | 20  |
| 1.3 | Histogram of inclusions data . . . . .  | 22  |
| 1.4 | Histogram for $Y_A$ data with superimposed theoretical distribution   | 24  |
| 1.5 | Histogram for $Y_B$ data with superimposed theoretical distribution   | 24  |
| 1.6 | Theoretical probability distribution function for a Poisson random variable with parameter $\lambda = 1.02$ . Compare with the inclusions data histogram in Fig 1.3 . . . . . | 25  |
| 2.1 | Schematic diagram of a plug flow reactor (PFR). . . . .   | 36  |
| 2.2 | Schematic diagram of a continuous stirred tank reactor (CSTR). . . . .  | 37  |
| 2.3 | Instantaneous residence time distribution function for the CSTR: (with $\tau = 5$ ). . . . .  | 39  |
| 3.1 | Venn Diagram for Example 3.7 . . . . .  | 66  |
| 3.2 | Venn diagram of students in a thermodynamics class . . . . .  | 72  |
| 3.3 | The role of “conditioning” Set B in conditional probability . . . . .   | 73  |
| 3.4 | Representing set A as a union of 2 disjoint sets . . . . .  | 74  |
| 3.5 | Partitioned sets for generalizing total probability result . . . . .  | 75  |
| 4.1 | The original sample space, $\Omega$ , and the corresponding space $V$ induced by the random variable $X$ . . . . .  | 91  |
| 4.2 | Probability distribution function, $f(x)$ , and cumulative distribution function, $F(x)$ , for 3-coin toss experiment of Example 4.1 . . . . .                                | 97  |
| 4.3 | Distribution of a negatively skewed random variable . . . . .   | 110 |
| 4.4 | Distribution of a positively skewed random variable . . . . .   | 110 |
| 4.5 | Distributions with reference kurtosis (solid line) and mild kurtosis (dashed line) . . . . .  | 111 |
| 4.6 | Distributions with reference kurtosis (solid line) and high kurtosis (dashed line) . . . . .  | 112 |
| 4.7 | The pdf of a continuous random variable $X$ with a mode at $x = 1$ . . . . .  | 117 |
| 4.8 | The cdf of a continuous random variable $X$ showing the lower and upper quartiles and the median . . . . .  | 118 |

|      |   |     |
|------|---|-----|
| 5.1  | Graph of the joint pdf for the 2-dimensional random variable of Example 5.5 . . . . .   | 149 |
| 5.2  | Positively correlated variables: $\rho = 0.923$ . . . . .   | 159 |
| 5.3  | Negatively correlated variables: $\rho = -0.689$ . . . . .  | 159 |
| 5.4  | Essentially uncorrelated variables: $\rho = 0.085$ . . . . .  | 160 |
| 6.1  | Region of interest, $V_Y$ , for computing the cdf of the random variable $Y$ defined as a sum of 2 independent random variables $X_1$ and $X_2$ . . . . .   | 178 |
| 6.2  | Schematic diagram of the tennis ball launcher of Problem 6.11 . . . . .   | 193 |
| 9.1  | Exponential pdfs for various values of parameter $\beta$ . . . . .  | 262 |
| 9.2  | Gamma pdfs for various values of parameter $\alpha$ and $\beta$ : Note how with increasing values of $\alpha$ the shape becomes less skewed, and how the breadth of the distribution increases with increasing values of $\beta$ . . . . .  | 267 |
| 9.3  | Gamma distribution fit to data on inter-origin distances in the budding yeast <i>S. cerevisiae</i> genome . . . . .   | 270 |
| 9.4  | Weibull pdfs for various values of parameter $\zeta$ and $\beta$ : Note how with increasing values of $\zeta$ the shape becomes less skewed, and how the breadth of the distribution increases with increasing values of $\beta$ . . . . .  | 274 |
| 9.5  | The Herschel-Maxwell 2-dimensional plane . . . . .  | 286 |
| 9.6  | Gaussian pdfs for various values of parameter $\mu$ and $\sigma$ : Note the symmetric shapes, how the center of the distribution is determined by $\mu$ , and how the shape becomes broader with increasing values of $\sigma$ . . . . .    | 289 |
| 9.7  | Symmetric tail area probabilities for the standard normal random variable with $z = \pm 1.96$ and $F_Z(-1.96) = 0.025 = 1 - F_Z(1.96)$ . . . . .  | 291 |
| 9.8  | Lognormal pdfs for scale parameter $\alpha = 0$ and various values of the shape parameter $\beta$ . Note how the shape changes, becoming less skewed as $\beta$ becomes smaller. . . . .  | 295 |
| 9.9  | Lognormal pdfs for shape parameter $\beta = 1$ and various values of the scale parameter $\alpha$ . Note how the shape remains unchanged while the entire distribution is scaled appropriately depending on the value of $\alpha$ . . . . . | 295 |
| 9.10 | Particle size distribution for the granulation process product: a log-normal distribution with $\alpha = 6.8, \beta = 0.5$ . The shaded area corresponds to product meeting quality specifications, $350 < X < 1650$ microns. . . . .       | 298 |
| 9.11 | Unimodal Beta pdfs when $\alpha > 1; \beta > 1$ : Note the symmetric shape when $\alpha = \beta$ , and the skewness determined by the value of $\alpha$ relative to $\beta$ . . . . .   | 304 |
| 9.12 | U-Shaped Beta pdfs when $\alpha < 1; \beta < 1$ . . . . .   | 304 |
| 9.13 | Other shapes of the Beta pdfs: It is J-shaped when $(\alpha-1)(\beta-1) < 0$ and a straight line when $\beta = 2; \alpha = 1$ . . . . .   | 305 |
| 9.14 | Theoretical distribution for characterizing fractional microarray intensities for the example gene: The shaded area corresponds to the probability that the gene in question is upregulated. . . . .  | 307 |

|  |     |
|--|-----|
| 9.15 Two uniform distributions over different ranges (0,1) and (2,10).<br>Since the total area under the pdf must be 1, the narrower pdf is<br>proportionately longer than the wider one. . . . .                      | 309 |
| 9.16 Two $F$ distribution plots for different values for $\nu_1$ , the first degree<br>of freedom, but the same value for $\nu_2$ . Note how the mode shifts to<br>the right as $\nu_1$ increases . . . . .            | 311 |
| 9.17 Three $t$ -distribution plots for degrees of freedom values $\nu =$<br>$5, 10, 100$ . Note the symmetrical shape and the heavier tail for<br>smaller values of $\nu$ . . . . .                                    | 312 |
| 9.18 A comparison of the $t$ -distribution with $\nu = 5$ with the standard<br>normal $N(0, 1)$ distribution. Note the similarity as well as the $t$ -<br>distribution's comparatively heavier tail. . . . .           | 313 |
| 9.19 A comparison of the $t$ -distribution with $\nu = 50$ with the standard<br>normal $N(0, 1)$ distribution. The two distributions are practically<br>indistinguishable. . . . .                                     | 313 |
| 9.20 A comparison of the standard Cauchy distributions with the stan-<br>dard normal $N(0, 1)$ distribution. Note the general similarities as<br>well as the Cauchy distribution's substantially heavier tail. . . . . | 315 |
| 9.21 Common probability distributions and connections among them . .   | 319 |
| 10.1 The entropy function of a Bernoulli random variable . . . . .   | 340 |
| 11.1 Elsner data versus binomial model prediction . . . . .  | 379 |
| 11.2 Elsner data ("Younger" set) versus binomial model prediction . .  | 381 |
| 11.3 Elsner data ("Older" set) versus binomial model prediction . . .  | 382 |
| 11.4 Elsner data ("Younger" set) versus stratified binomial model predic-<br>tion . . . . .  | 383 |
| 11.5 Elsner data ("Older" set) versus stratified binomial model prediction   | 383 |
| 11.6 Complete Elsner data versus stratified binomial model prediction .  | 384 |
| 11.7 Optimum number of embryos as a function of $p$ . . . . .  | 386 |
| 11.8 Surface plot of the probability of a singleton as a function of $p$ and<br>the number of embryos transferred, $n$ . . . . .   | 388 |
| 11.9 The (maximized) probability of a singleton as a function of $p$ when<br>the optimum integer number of embryos are transferred . . . . .   | 388 |
| 11.10 Surface plot of the probability of no live birth as a function of $p$ and<br>the number of embryos transferred, $n$ . . . . .  | 389 |
| 11.11 Surface plot of the probability of multiple births as a function of $p$<br>and the number of embryos transferred, $n$ . . . . .  | 389 |
| 11.12 IVF treatment outcome probabilities for "good prognosis" patients<br>with $p = 0.5$ , as a function of $n$ , the number of embryos transferred   | 391 |
| 11.13 IVF treatment outcome probabilities for "medium prognosis" pa-<br>tients with $p = 0.3$ , as a function of $n$ , the number of embryos trans-<br>ferred . . . . .  | 392 |
| 11.14 IVF treatment outcome probabilities for "poor prognosis" patients<br>with $p = 0.18$ , as a function of $n$ , the number of embryos transferred  | 393 |

|   |     |
|---|-----|
| 11.15Relative sensitivity of the binomial model derived $n^*$ to errors in estimates of $p$ as a function of $p$ . . . . .  | 396 |
| 12.1 Relating the tools of Probability, Statistics and Design of Experiments to the concepts of Population and Sample . . . . .   | 415 |
| 12.2 Bar chart of welding injuries from Table 12.1 . . . . .  | 420 |
| 12.3 Bar chart of welding injuries arranged in decreasing order of number of injuries . . . . .   | 420 |
| 12.4 Pareto chart of welding injuries . . . . .   | 421 |
| 12.5 Pie chart of welding injuries . . . . .  | 422 |
| 12.6 Bar Chart of frozen ready meals sold in France in 2002 . . . . .   | 423 |
| 12.7 Pie Chart of frozen ready meals sold in France in 2002 . . . . .   | 424 |
| 12.8 Histogram for $Y_A$ data of Chapter 1 . . . . .  | 425 |
| 12.9 Frequency Polygon of $Y_A$ data of Chapter 1 . . . . .   | 427 |
| 12.10Frequency Polygon of $Y_B$ data of Chapter 1 . . . . .   | 428 |
| 12.11Boxplot of the chemical process yield data $Y_A$ , $Y_B$ of Chapter 1 . .  | 429 |
| 12.12Boxplot of random $N(0,1)$ data: original set, and with added “outlier” . . . . .  | 430 |
| 12.13Box plot of raisins dispensed by five different machines . . . . .   | 431 |
| 12.14Scatter plot of cranial circumference versus finger length: The plot shows no real relationship between these variables . . . . .  | 432 |
| 12.15Scatter plot of city gas mileage versus highway gas mileage for various two-seater automobiles: The plot shows a strong positive linear relationship, but no causality is implied. . . . .   | 433 |
| 12.16Scatter plot of highway gas mileage versus engine capacity for various two-seater automobiles: The plot shows a negative linear relationship. Note the two unusually high mileage values associated with engine capacities 7.0 and 8.4 liters identified as belonging to the Chevrolet Corvette and the Dodge Viper, respectively. . . . . | 434 |
| 12.17Scatter plot of highway gas mileage versus number of cylinders for various two-seater automobiles: The plot shows a negative linear relationship. . . . .  | 434 |
| 12.18Scatter plot of US population every ten years since the 1790 census versus census year: The plot shows a strong non-linear trend, with very little scatter, indicative of the systematic, approximately exponential growth . . . . .   | 435 |
| 12.19Scatter plot of $Y_1$ and $X_1$ from Anscombe data set 1. . . . .  | 444 |
| 12.20Scatter plot of $Y_2$ and $X_2$ from Anscombe data set 2. . . . .  | 445 |
| 12.21Scatter plot of $Y_3$ and $X_3$ from Anscombe data set 3. . . . .  | 445 |
| 12.22Scatter plot of $Y_4$ and $X_4$ from Anscombe data set 4. . . . .  | 446 |
| 13.1 Sampling distribution for mean lifetime of DLP lamps in Example 13.3 used to compute $P(5100 < \bar{X} < 5200) = P(-0.66 < Z < 1.34)$ . . . . .  | 469 |
| 13.2 Sampling distribution for average lifetime of DLP lamps in Example 13.3 used to compute $P(\bar{X} < 5000) = P(Z < -2.67)$ . . . . .   | 470 |

|   |     |
|---|-----|
| 13.3 Sampling distribution of the mean diameter of ball bearings in Example 13.4 used to compute $P( \bar{X} - 10  \geq 0.14) = P( T  \geq 0.62)$ . . . . .   | 473 |
| 13.4 Sampling distribution for the variance of ball bearing diameters in Example 13.5 used to compute $P(S \geq 1.01) = P(C \geq 23.93)$ . . . . .  | 475 |
| 13.5 Sampling distribution for the two variances of ball bearing diameters in Example 13.6 used to compute $P(F \geq 1.41) + P(F \leq 0.709)$ . . . . .   | 476 |
| <br>  |     |
| 14.1 Sampling distribution for the two estimators $U_1$ and $U_2$ : $U_1$ is the more efficient estimator because of its smaller variance . . . . .   | 491 |
| 14.2 Two-sided tail area probabilities of $\alpha/2$ for the standard normal sampling distribution . . . . .  | 504 |
| 14.3 Two-sided tail area probabilities of $\alpha/2 = 0.025$ for a Chi-squared distribution with 9 degrees of freedom . . . . .   | 511 |
| 14.4 Sampling distribution with two-sided tail area probabilities of 0.025 for $\bar{X}/\beta$ , based on a sample of size $n = 10$ from an exponential population . . . . .  | 516 |
| 14.5 Sampling distribution with two-sided tail area probabilities of 0.025 for $\bar{X}/\beta$ , based on a <i>larger</i> sample of size $n = 100$ from an exponential population . . . . .   | 517 |
| <br>  |     |
| 15.1 A distribution for the null hypothesis, $H_0$ , in terms of the test statistic, $Q_T$ , where the shaded rejection region, $Q_T > q$ , indicates a significance level, $\alpha$ . . . . .  | 557 |
| 15.2 Overlapping distributions for the null hypothesis, $H_0$ (with mean $\mu_0$ ), and alternative hypothesis, $H_a$ (with mean $\mu_a$ ), showing Type I and Type II error risks $\alpha, \beta$ , along with $q_C$ the boundary of the critical region of the test statistic, $Q_T$ . . . . .  | 559 |
| 15.3 The standard normal variate $z = -z_\alpha$ with tail area probability $\alpha$ . The shaded portion is the rejection region for a lower-tailed test, $H_a : \mu < \mu_0$ . . . . .  | 564 |
| 15.4 The standard normal variate $z = z_\alpha$ with tail area probability $\alpha$ . The shaded portion is the rejection region for an upper-tailed test, $H_a : \mu > \mu_0$ . . . . .  | 565 |
| 15.5 Symmetric standard normal variates $z = z_{\alpha/2}$ and $z = -z_{\alpha/2}$ with identical tail area probabilities $\alpha/2$ . The shaded portions show the rejection regions for a two-sided test, $H_a : \mu \neq \mu_0$ . . . . .  | 565 |
| 15.6 Box plot for Method A scores including the null hypothesis mean, $H_0 : \mu = 75$ , shown along with the sample average, $\bar{x}$ , and the 95% confidence interval based on the $t$ -distribution with 9 degrees of freedom. Note how the upper bound of the 95% confidence interval lies to the left of, and does not touch, the postulated $H_0$ value . . . . . | 574 |

|   |     |
|---|-----|
| 15.7 Box plot for Method B scores including the null hypothesis mean, $H_0, \mu = 75$ , shown along with the sample average, $\bar{x}$ , and the 95% confidence interval based on the $t$ -distribution with 9 degrees of freedom. Note how the the 95% confidence interval includes the postulated $H_0$ value . . . . .   | 574 |
| 15.8 Box plot of differences between the “Before” and “After” weights, including a 95% confidence interval for the mean difference, and the hypothesized $H_0$ point, $\delta_0 = 0$ . . . . .  | 588 |
| 15.9 Box plot of the “Before” and “After” weights including individual data means. Notice the wide range of each data set . . . . .   | 590 |
| 15.10A plot of the “Before” and “After” weights for each patient. Note how one data sequence is almost perfectly correlated with the other; in addition note the relatively large variability intrinsic in each data set compared to the difference between each point . . . . .  | 590 |
| 15.11 Null and alternative hypotheses distributions for upper-tailed test based on $n = 25$ observations, with population standard deviation $\sigma = 4$ , where the true alternative mean, $\mu_a$ , exceeds the hypothesized one by $\delta^* = 2.0$ . The figure shows a “z-shift” of $(\delta^* \sqrt{n})/\sigma = 2.5$ ; and with reference to $H_0$ , the critical value $z_{0.05} = 1.65$ . The area under the $H_0$ curve to the <i>right</i> of the point $z = 1.65$ is $\alpha = 0.05$ , the significance level; the area under the dashed $H_a$ curve to the <i>left</i> of the point $z = 1.65$ is $\beta$ . . . . . | 592 |
| 15.12 $\beta$ and power values for hypothesis test of Fig 15.11 with $H_a \sim N(2.5, 1)$ . Top: $\beta$ ; Bottom: Power = $(1 - \beta)$ . . . . .  | 594 |
| 15.13 Rejection regions for one-sided tests of a single variance of a normal population, at a significance level of $\alpha = 0.05$ , based on $n = 10$ samples. The distribution is $\chi^2(9)$ ; Top: for $H_a : \sigma^2 > \sigma_0^2$ , indicating rejection of $H_0$ if $c^2 > \chi_{\alpha}^2(9) = 16.9$ ; Bottom: for $H_a : \sigma^2 < \sigma_0^2$ , indicating rejection of $H_0$ if $c^2 < \chi_{1-\alpha}^2(9) = 3.33$ . . . . .   | 602 |
| 15.14 Rejection regions for the two-sided tests concerning the variance of the process A yield data $H_0 : \sigma_A^2 = 1.5^2$ , based on $n = 50$ samples, at a significance level of $\alpha = 0.05$ . The distribution is $\chi^2(49)$ , with the rejection region shaded; because the test statistic, $c^2 = 44.63$ , falls outside of the rejection region, we do not reject $H_0$ . . . . .   | 604 |
| 15.15 Rejection regions for the two-sided tests of the equality of the variances of the process A and process B yield data, i.e., $H_0 : \sigma_A^2 = \sigma_B^2$ , at a significance level of $\alpha = 0.05$ , based on $n = 50$ samples each. The distribution is $F(49, 49)$ , with the rejection region shaded; since the test statistic, $f = 0.27$ , falls within the rejection region to the left, we reject $H_0$ in favor of $H_a$ . . . . .  | 606 |
| 16.1 Boiling point of hydrocarbons in Table 16.1 as a function of the number of carbon atoms in the compound . . . . .  | 649 |
| 16.2 The true regression line and the zero mean random error $\epsilon_i$ . . . . .   | 654 |

|   |     |
|---|-----|
| 16.3 The Gaussian assumption regarding variability around the true regression line giving rise to $\epsilon \sim N(0, \sigma^2)$ : The 6 points represent the data at $x_1, x_2, \dots, x_6$ ; the solid straight line is the true regression line which passes through the mean of the sequence of the indicated Gaussian distributions . . . . .  | 655 |
| 16.4 The fitted straight line to the Density versus Ethanol Weight % data: The additional terms included in the graph, $S$ , $R\text{-Sq}$ and $R\text{-Sq(adj)}$ are discussed later . . . . .   | 659 |
| 16.5 The fitted regression line to the Density versus Ethanol Weight % data (solid line) along with the 95% confidence interval (dashed line). The confidence interval is narrowest at $x = \bar{x}$ and widens for values further away from $\bar{x}$ . . . . .  | 664 |
| 16.6 The fitted straight line to the Cranial circumference versus Finger length data. Note how the data points are widely scattered around the fitted regression line. (The additional terms included in the graph, $S$ , $R\text{-Sq}$ and $R\text{-Sq(adj)}$ are discussed later) . . . . .   | 667 |
| 16.7 The fitted straight line to the Highway MPG versus Engine Capacity data of Table 12.5 (leaving out the two “inconsistent” data points) along with the 95% confidence interval (long dashed line) and the 95% prediction interval (short dashed line). (Again, the additional terms included in the graph, $S$ , $R\text{-Sq}$ and $R\text{-Sq(adj)}$ are discussed later). . . . .   | 670 |
| 16.8 Modeling the temperature dependence of thermal conductivity: Top: Fitted straight line to the Thermal conductivity ( $k$ ) versus Temperature ( $T^\circ C$ ) data in Table 16.6; Bottom: standardized residuals versus fitted value, $\hat{y}_i$ . . . . .  | 681 |
| 16.9 Modeling the dependence of the boiling points (BP) of hydrocarbon compounds in Table 16.1 on the number of carbon atoms in the compound: Top: Fitted straight line of BP versus $n$ , the number of carbon atoms; Bottom: standardized residuals versus fitted value $\hat{y}_i$ . Notice the distinctive quadratic structure “left over” in the residuals exposing the linear model’s over-estimation at the extremes and the under-estimation in the middle. . . . . | 683 |
| 16.10 Catalytic process yield data of Table 16.7 . . . . .  | 692 |
| 16.11 Catalytic process yield data of Table 16.1. Top: Fitted plane of Yield as a function of Temperature and Pressure; Bottom: standardized residuals versus fitted value $\hat{y}_i$ . Nothing appears unusual about these residuals. . . . .   | 695 |
| 16.12 Modeling the dependence of the boiling points (BP) of hydrocarbon compounds in Table 16.1 on the number of carbon atoms in the compound: Top: Fitted quadratic curve of BP versus $n$ , the number of carbon atoms; Bottom: standardized residuals versus fitted value $\hat{y}_i$ . Despite the good fit, the visible systematic structure still “left over” in the residuals suggests adding one more term to the model.  | 703 |

|  |     |
|--|-----|
| 16.13 Modeling the dependence of the boiling points (BP) of hydrocarbon compounds in Table 16.1 on the number of carbon atoms in the compound: Top: Fitted cubic curve of BP versus $n$ , the number of carbon atoms; Bottom: standardized residuals versus fitted value $\hat{y}_i$ . There appears to be little or no systematic structure left in the residuals, suggesting that the cubic model provides an adequate description of the data. . . . .                                | 705 |
| 16.14 Gram polynomials evaluated at 5 discrete points $k = 1, 2, 3, 4, 5$ ; $p_0$ is the constant; $p_1$ , the straight line; $p_2$ , the quadratic and $p_3$ , the cubic . . . . .  | 707 |
| 17.1 Probability plots for safety data postulated to be exponentially distributed, each showing (a) rank ordered data; (b) theoretical fitted cumulative probability distribution line along with associated 95% confidence intervals; (c) a list of summary statistics, including the $p$ -value associated with a formal goodness-of-fit test. The indication from the $p$ -values is that there is no evidence to reject $H_0$ ; therefore the model appears to be adequate . . . . . | 738 |
| 17.2 Probability plot for safety data $S_2$ wrongly postulated to be normally distributed. The departure from the linear fit does not appear too severe, but the low/borderline $p$ -value (0.045) objectively compels us to reject $H_0$ at the 0.05 significance level and conclude that the Gaussian model is inadequate for this data. . . . .   | 739 |
| 17.3 Probability plots for yield data sets $Y_A$ and $Y_B$ postulated to be normally distributed. The 95% confidence intervals around the fitted line, along with the indicated $p$ -values, strongly suggest that the distributional assumptions appear to be valid. . . . .  | 740 |
| 17.4 Normal probability plot for the residuals of the regression analysis of the dependence of thermal conductivity, $k$ , on Temperature in Example 16.5. The postulated model, a two-parameter regression model with Gaussian distributed zero mean errors, appears valid. . . . .   | 741 |
| 17.5 Chi-Squared test results for inclusions data and a postulated Poisson model. Top panel: Bar chart of “Expected” and “Observed” frequencies, which shows how well the model prediction matches observed data; Bottom Panel: Bar chart of contributions to the Chi-squared statistic, showing that the group of 3 or more inclusions is responsible for the largest model-observation discrepancy, by a wide margin. . . . .  | 744 |
| 18.1 Histograms of interspike intervals data with Gamma model fit for the pyramidal tract cell of a monkey. Top panel: when awake (PT-W); Bottom Panel: when asleep (PT-S). Note the <i>similarities</i> in the estimated values for $\alpha$ —the shape parameter—for both sets of data, and the <i>difference</i> between the estimates for $\beta$ , the scale parameters. . . . .  | 774 |

|   |     |
|---|-----|
| 18.2 Probability plot of interspike intervals data with postulated Gamma model and Anderson-Darling test for the pyramidal tract cell of a monkey. Top panel: when awake (PT-W); Bottom panel: when asleep (PT-S). The $p$ -values for the A-D tests indicate no evidence to reject the null hypothesis . . . . .   | 776 |
| 19.1 Graphic illustration of the orthogonal vector decomposition of Eq (19.11) . . . . .  | 800 |
| 19.2 Boxplot of raisins data showing what the ANOVA analysis has confirmed that there is a significant difference in how the machines dispense raisins. . . . .   | 802 |
| 19.3 Normal probability plots of the residuals from the one-way classification ANOVA model in Example 19.1. Top panel: Plot obtained directly from the ANOVA analysis which does not provide any test statistic or significance level; Bottom panel: Subsequent goodness-of-fit test carried out on saved residuals; note the high $p$ -value associated with the A-D test. . . . .   | 804 |
| 19.4 Graphic illustration of the orthogonal error decomposition of Eq (19.21) with the additional block component, $E_B$ . . . . .  | 807 |
| 19.5 Normal probability plots of the residuals from the two-way classification ANOVA model for investigating tire wear, obtained directly from the ANOVA analysis. . . . .  | 810 |
| 19.6 $2^2$ factorial design for factors $A$ and $B$ showing the four experimental points; – represents low values, + represents high values for each factor. . . . .  | 815 |
| 19.7 Graphic illustration of “folding” where two half-fractions of a $2^3$ factorial design are combined to recover the full factorial design; each fold costs an additional degree of freedom for analysis. . . . .  | 826 |
| 19.8 Normal probability plot for the effects, using Lenth’s method to identify $A$ , $D$ and $AD$ as significant. . . . .   | 830 |
| 19.9 Normal probability plot for the residuals of the Etch rate model in Eq (19.46) obtained upon projection of the experimental data to retain only the significant terms $A$ , Gap ( $x_1$ ), $D$ , Power ( $x_2$ ), and the interaction $AD$ , Gap*Power ( $x_1x_2$ ). . . . .   | 832 |
| 19.10 The 3-factor face-centered cube (FCC) response surface design and its constituent parts: $2^3$ factorial base, Open circles; face center points, lighter shaded circles; center point, darker solid circle. . . . .   | 835 |
| 19.11 The 3-factor Box-Behnken response surface design and its constituent parts: $X_1, X_2$ : $2^2$ factorial points moved to the center of $X_3$ to give the darker shaded circles at the edge-centers of the $X_3$ axes; $X_2, X_3$ : $2^2$ factorial points moved to the center of $X_1$ to give the lighter shaded circles at the edge-centers of the $X_1$ axes; $X_1, X_3$ : $2^2$ factorial points moved to the center of $X_2$ to give the solid circles at the edge-centers of the $X_2$ axes; the center point, open circle. . . . . | 836 |

|       |   |     |
|-------|---|-----|
| 20.1  | Chi-Squared test results for Prussian army death by horse kicks data and a postulated Poisson model. Top panel: Bar chart of “Expected” and “Observed” frequencies; Bottom Panel: Bar chart of contributions to the Chi-squared statistic. . . . .  | 861 |
| 20.2  | Initial prior distribution, a Gamma (2,0.5), used to obtain a Bayesian estimate for the Poisson mean number of deaths per unit-year parameter. . . . .  | 864 |
| 20.3  | Recursive Bayesian estimates using yearly data sequentially, compared with the standard maximum likelihood estimate, 0.61, (dashed-line). . . . .   | 867 |
| 20.4  | Final posterior distribution (dashed line) along with initial prior distribution (solid line). . . . .  | 868 |
| 20.5  | Quadratic regression model fit to US Population data along with both the 95% confidence interval and the 95% prediction interval. . . . .   | 874 |
| 20.6  | Standardized residuals from the regression model fit to US Population data. Top panel: Residuals versus observation order; Bottom panel: Normal probability plot. Note the left-over pattern indicative of serial correlation, and the “unusual” observations identified for the 1940 and 1950 census years in the top panel; note also the general deviation of the residuals from the theoretical normal probability distribution line in the bottom panel. . . . . | 875 |
| 20.7  | Percent average relative population growth rate in the US for each census year from 1800-2000 divided into three equal 70-year periods. Period 1: 1800-1860; Period 2: 1870-1930; Period 3: 1940-2000. . .  | 877 |
| 20.8  | Normal probability plot for the residuals from the ANOVA model for Percent average relative population growth rate versus Period with Period 1: 1800-1860; Period 2: 1870-1930; Period 3: 1940-2000. .  | 878 |
| 20.9  | Standardized residual plots for “Yield” response surface model: versus fitted value, and normal probability plot. . . . .   | 884 |
| 20.10 | Standardized residual plots for “Adhesion” response surface model: versus fitted value, and normal probability plot. . . . .  | 885 |
| 20.11 | Response surface and contour plots for “Yield” as a function of Additive and Temperature (with Time held at 60.00). . . . .   | 886 |
| 20.12 | Response surface and contour plots for “Adhesion” as a function of Additive and Temperature (with Time held at 60.00). . . . .  | 887 |
| 20.13 | Overlaid contours for “Yield” and “Adhesion” showing feasible region for desired optimum. The planted “flag” indicates the optimum values of the responses along with the corresponding setting of the factors Additive and Temperature (with Time held at 60.00) that achieve this optimum. . . . .  | 888 |
| 20.14 | Schematic diagram of folded helicopter prototype . . . . .  | 891 |
| 20.15 | Paper helicopter prototype . . . . .  | 893 |
| 21.1  | Simple Systems: Series and parallel configuration . . . . .   | 902 |
| 21.2  | A series-parallel arrangement of a 6-component system . . . . .   | 902 |

|  |     |
|--|-----|
| 21.3 Sampling-analyzer system: basic configuration . . . . .   | 907 |
| 21.4 Sampling-analyzer system: configuration with redundant solenoid valve . . . . .   | 907 |
| 21.5 Fluid flow system with a cross link . . . . .   | 909 |
| 21.6 Typical failure rate (hazard function) curve showing the classic three distinct characteristic periods in the lifetime distributions of a population of items . . . . .   | 913 |
| 21.7 Blood storage system . . . . .  | 926 |
| 21.8 Nuclear power plant heat exchanger system . . . . .   | 927 |
| 21.9 Fluid flow system with a cross link (from Fig 21.5) . . . . .   | 927 |
| 21.10 Fire alarm system with back up . . . . .   | 928 |
| 21.11 Condenser system for VOCs . . . . .  | 928 |
| 21.12 Simplified representation of the control structure in the baroreceptor reflex . . . . .  | 929 |
| 22.1 OC Curve for a lot size of 1000, sample size of 32 and acceptance number of 3: AQL is the acceptance quality level; RQL is the rejection quality level. . . . .   | 939 |
| 22.2 OC Curve for a lot size of 1000, generated for a sampling plan for an AQL = 0.004 and an RQL = 0.02, leading to a required sample size of 333 and acceptance number of 3. Compare with the OC curve in Fig 22.1. . . . .                                | 943 |
| 22.3 A generic SPC chart for the generic process variable $Y$ indicating a sixth data point that is out of limits. . . . .   | 946 |
| 22.4 The X-bar chart for the average length measurements for 6-inch nails determined from samples of three measurements obtained every 5 mins. . . . .   | 948 |
| 22.5 The S-chart for the 6-inch nails process data of Example 22.2. . .  | 951 |
| 22.6 The combination Xbar-R chart for the 6-inch nails process data of Example 22.2. . . . .   | 952 |
| 22.7 The combination I-MR chart for the Mooney viscosity data. . . . .   | 954 |
| 22.8 P-chart for the data on defective mechanical pencils: note the 9 <sup>th</sup> observation that is outside the UCL. . . . .   | 956 |
| 22.9 C-chart for the inclusions data presented in Chapter 1, Table 1.2, and discussed in subsequent chapters: note the 33 <sup>rd</sup> observation that is outside the UCL, otherwise, the process appears to be operating in statistical control . . . . . | 958 |
| 22.10 Time series plot of the original Mooney viscosity data of Fig 22.7 and Table 22.2, and of the shifted version showing a step increase of 0.7 after sample 15. . . . .  | 959 |
| 22.11 I-chart for the shifted Mooney viscosity data. Even with $\sigma = 0.5$ , it is not sensitive enough to detect the step change of 0.7 introduced after sample 15. . . . .  | 960 |

|       |   |      |
|-------|---|------|
| 22.12 | Two one-sided CUSUM charts for the shifted Mooney viscosity data.<br>The upper chart uses dots; the lower chart uses diamonds; the non-conforming points are represented with the squares. With the same $\sigma = 0.5$ , the step change of 0.7 introduced after sample 15 is identified after sample 18. Compare with the I-Chart in Fig 22.11. . . . .   | 962  |
| 22.13 | Two one-sided CUSUM charts for the original Mooney viscosity data using the same characteristics as those in Fig 22.12. The upper chart uses dots; the lower chart uses diamonds; there are no non-conforming points. . . . .   | 962  |
| 22.14 | EWMA chart for the shifted Mooney viscosity data, with $w = 0.2$ .<br>Note the staircase shape of the control limits for the earlier data points. With the same $\sigma = 0.5$ , the step change of 0.7 introduced after sample 15 is detected after sample 18. The non-conforming points are represented with the squares. Compare with the I-Chart in Fig 22.11 and the CUSUM charts in Fig 22.12. . . . .  | 964  |
| 22.15 | The EWMA chart for the original Mooney viscosity data using the same characteristics as in Fig 22.14. There are no non-conforming points. . . . .   | 965  |
| 23.1  | Examples of the bivariate Gaussian distribution where the two random variables are uncorrelated ( $\rho = 0$ ) and strongly positively correlated ( $\rho = 0.9$ ). . . . .   | 981  |
| 23.2  | Plot of the 16 variables in the illustrative example data set. . . . .  | 992  |
| 23.3  | Scree plot showing that the first two components are the most important. . . . .  | 994  |
| 23.4  | Plot of the scores and loading for the first principal component. The distinct trend indicated in the scores should be interpreted along with the loadings by comparison to the full original data set in Fig 23.2. . . . .   | 995  |
| 23.5  | Plot of the scores and loading for the second principal component. The distinct trend indicated in the scores should be interpreted along with the loadings by comparison to the full original data set in Fig 23.2. . . . .  | 996  |
| 23.6  | Scores and loading plots for the first two components. Top panel: Scores plot indicates a quadratic relationship between the two scores $t_1$ and $t_2$ ; Bottom panel: Loading vector plot indicates that in the new set of coordinates, the original variables contain mostly pure components PC1 and PC2 indicated by a distinctive North/South and West/East alignment of the data vectors, with like variables clustered together according to the nature of the component contributions. Compare to the full original data set in Fig 23.2. . . . . | 998  |
| 23.7  | Principal component model for a 3-dimensional data set described by two principal components on a plane, showing a point with a large $Q$ and another with a large $T^2$ value. . . . .   | 1001 |

|   |      |
|---|------|
| 23.8 Control limits for $Q$ and $T^2$ for process data represented with two principal components. . . . . | 1001 |
|---|------|



---

## ***List of Tables***

|     |   |     |
|-----|---|-----|
| 1.1 | Yield Data for Process A versus Process B . . . . .   | 13  |
| 1.2 | Number of “inclusions” on sixty 1-sq meter glass sheets . . . . .   | 16  |
| 1.3 | Group classification and frequencies for $Y_A$ data . . . . .   | 18  |
| 1.4 | Group classification and frequencies for $Y_B$ data . . . . .   | 19  |
| 1.5 | Group classification and frequencies for the <i>inclusions</i> data . .   | 21  |
| 2.1 | Computed probabilities of occurrence of various number of <i>inclusions</i> for $\lambda = 2$ . . . . .                 | 44  |
| 3.1 | Subsets and Events . . . . .  | 63  |
| 3.2 | Class list and attributes . . . . .   | 65  |
| 3.3 | Lithium toxicity study results . . . . .  | 85  |
| 4.1 | $f(x)$ and $F(x)$ for the “three coin-toss” experiments of Example 4.1 . . . . .  | 96  |
| 4.2 | The pdf $f(x)$ for the ball-drawing game . . . . .  | 103 |
| 4.3 | Summary analysis for the ball-drawing game . . . . .  | 104 |
| 5.1 | Joint pdf for computer store sales . . . . .  | 151 |
| 5.2 | Joint and marginal pdfs for computer store sales . . . . .  | 152 |
| 5.3 | Conditional pdf $f(x_1 x_2)$ for computer store sales . . . . .   | 152 |
| 5.4 | Conditional pdf $f(x_2 x_1)$ for computer store sales . . . . .   | 153 |
| 5.5 | Joint and marginal pdfs for two-coin toss problem of Example 5.1 . . . . .  | 162 |
| 7.1 | Summary of Mendel’s single trait experiment results . . . . .   | 202 |
| 7.2 | Theoretical distribution of shape-color traits in second generation hybrids under the independence assumption . . . . . | 207 |
| 7.3 | Theoretical versus experimental results for second generation hybrid plants . . . . .                                   | 208 |
| 7.4 | Attacks and hits on US Naval Warships in 1943 . . . . .   | 210 |
| 8.1 | Theoretical versus empirical frequencies for <i>inclusions</i> data . .   | 241 |
| 8.2 | Summary of probability models for discrete random variables   | 245 |
| 9.1 | Summary of probability models for continuous random variables . . . . .   | 318 |

|   |     |
|---|-----|
| 10.1 Summary of maximum entropy probability models . . . . .  | 356 |
| 11.1 Theoretical distribution of probabilities of possible outcomes of an IVF treatment . . . . .                     | 373 |
| 11.2 Elsner, <i>et al.</i> , data of outcomes of a 42-month IVF treatment study . . . . .                             | 376 |
| 11.3 Binomial model prediction of Elsner, <i>et al.</i> data in Table 11.2  | 378 |
| 11.4 Elsner data stratified by age indicating variability in the “probability of success” estimates . . . . .         | 379 |
| 11.5 Stratified binomial model prediction of Elsner, <i>et al.</i> data. . .  | 382 |
| 12.1 Number and Type of injuries incurred by welders in the USA from 1980-1989 . . . . .                              | 419 |
| 12.2 Frozen Ready meals in France, in 2002 . . . . .  | 422 |
| 12.3 Group classification and frequencies for $Y_A$ data . . . . .  | 425 |
| 12.4 Number of raisins dispensed into trial-sized “Raisin Bran” cereal boxes . . . . .                                | 430 |
| 12.5 Gasoline mileage ratings for a collection of two-seater automobiles . . . . .                                    | 433 |
| 12.6 Descriptive statistics for yield data sets $Y_A$ and $Y_B$ . . . . .   | 441 |
| 12.7 The Anscombe data set 1 . . . . .  | 443 |
| 12.8 The Anscombe data sets 2, 3, and 4 . . . . .   | 443 |
| 14.1 Summary of estimation results . . . . .  | 549 |
| 14.2 Some population parameters and conjugate prior distributions appropriate for their Bayesian estimation . . . . . | 550 |
| 15.1 Hypothesis test decisions and risks . . . . .  | 558 |
| 15.2 Summary of $H_0$ rejection conditions for the one-sample $z$ -test   | 566 |
| 15.3 Summary of $H_0$ rejection conditions for the one-sample $t$ -test   | 571 |
| 15.4 Summary of $H_0$ rejection conditions for the two-sample $z$ -test   | 577 |
| 15.5 Summary of $H_0$ rejection conditions for the two-sample $t$ -test   | 579 |
| 15.6 “Before” and “After” weights for patients on a supervised weight-loss program . . . . .                          | 586 |
| 15.7 Summary of $H_0$ rejection conditions for the paired $t$ -test . .   | 587 |
| 15.8 Sample size $n$ required to achieve a power of 0.9 . . . . .   | 598 |
| 15.9 Summary of $H_0$ rejection conditions for the $\chi^2$ -test . . . . .   | 601 |
| 15.10 Summary of $H_0$ rejection conditions for the $F$ -test . . . . .   | 604 |
| 15.11 Summary of $H_0$ rejection conditions for the single-proportion $z$ -test . . . . .                             | 608 |
| 15.12 Summary of Selected Hypothesis Tests and their Characteristics . . . . .  | 645 |
| 16.1 Boiling points of a series of hydrocarbons . . . . .   | 649 |
| 16.2 Density (in gm/cc) and weight percent of ethanol in ethanol-water mixture . . . . .                              | 658 |

|   |     |
|---|-----|
| 16.3 Density and weight percent of ethanol in ethanol-water mixture: model fit and residual errors . . . . .                            | 659 |
| 16.4 Cranial circumference and finger lengths . . . . .   | 666 |
| 16.5 ANOVA Table for Testing Significance of Regression . . . . .   | 675 |
| 16.6 Thermal conductivity measurements at various temperatures for a metal . . . . .  | 679 |
| 16.7 Laboratory experimental data on Yield . . . . .  | 693 |
| 17.1 Table of values for safety data probability plot . . . . .   | 735 |
| 18.1 A professor's teaching evaluation scores organized by student type . . . . .   | 759 |
| 18.2 Interspike intervals data . . . . .  | 773 |
| 18.3 Summary of Selected Nonparametric Tests and their Characteristics . . . . .  | 779 |
| 19.1 Data table for typical single-factor experiment . . . . .  | 799 |
| 19.2 One-Way Classification ANOVA Table . . . . .   | 801 |
| 19.3 Data table for typical single-factor, two-way classification, experiment . . . . .   | 806 |
| 19.4 Two-Way Classification ANOVA Table . . . . .   | 808 |
| 19.5 Data table for typical two-factor experiment . . . . .   | 813 |
| 19.6 Two-factor ANOVA Table . . . . .   | 813 |
| 20.1 Frequency distribution of Prussian army deaths by horse kicks  | 858 |
| 20.2 Actual vs Predicted Frequency distribution of Prussian army deaths . . . . .   | 859 |
| 20.3 Year-by-Year, Unit-by-Unit breakdown of Prussian army deaths data . . . . .  | 862 |
| 20.4 Recursive (yearly) Bayesian estimates of the mean number of deaths per unit-year . . . . .   | 866 |
| 20.5 Frequency distribution of bomb hits in greater London during WW II and Poisson model prediction . . . . .                          | 869 |
| 20.6 US Population (to the nearest million) from 1790–2000 . . . . .  | 871 |
| 20.7 Percent average relative population growth rate for each census year . . . . .   | 877 |
| 20.8 Response surface design and experimental results for coating process . . . . .   | 880 |
| 21.1 Summary of $H_0$ rejection conditions for the test of hypothesis based on an exponential model of component failure-time . . . . . | 921 |
| 22.1 Measured length of samples of 6-inch nails in a manufacturing process . . . . .  | 949 |
| 22.2 Hourly Mooney viscosity data . . . . .   | 953 |
| 22.3 Number and proportion of defective mechanical pencils . . . . .  | 956 |



---

# **Contents**

|  |           |
|--|-----------|
| <b>0 Prelude</b>   | <b>1</b>  |
| 0.1 Approach Philosophy . . . . .                                      | 1         |
| 0.2 Four basic principles . . . . .                                    | 3         |
| 0.3 Summary and Conclusions . . . . .                                  | 5         |
| <b>I Foundations</b>   | <b>7</b>  |
| <b>1 Two Motivating Examples</b>                                       | <b>11</b> |
| 1.1 Yield Improvement in a Chemical Process . . . . .                  | 12        |
| 1.1.1 The Problem . . . . .  | 12        |
| 1.1.2 The Essence of the Problem . . . . .                             | 14        |
| 1.1.3 Preliminary “Intuitive” Notions . . . . .                        | 14        |
| 1.2 Quality Assurance in a Glass Sheet Manufacturing Process . . . . . | 16        |
| 1.3 Outline of a Systematic Approach . . . . .                         | 17        |
| 1.3.1 Group Classification and Frequency Distributions . . . . .       | 18        |
| 1.3.2 Theoretical Distributions . . . . .                              | 22        |
| 1.4 Summary and Conclusions . . . . .                                  | 25        |
| <b>2 Random Phenomena, Variability and Uncertainty</b>                 | <b>33</b> |
| 2.1 Two Extreme Idealizations of Natural Phenomena . . . . .           | 34        |
| 2.1.1 Introduction . . . . .   | 34        |
| 2.1.2 A Chemical Engineering Illustration . . . . .                    | 35        |
| 2.2 Random Mass Phenomena . . . . .                                    | 41        |
| 2.2.1 Defining Characteristics . . . . .                               | 41        |
| 2.2.2 Variability and Uncertainty . . . . .                            | 42        |
| 2.2.3 Practical Problems of Interest . . . . .                         | 42        |
| 2.3 Introducing Probability . . . . .                                  | 43        |
| 2.3.1 Basic Concepts . . . . .   | 43        |
| 2.3.2 Interpreting Probability . . . . .                               | 44        |
| 2.4 The Probabilistic Framework . . . . .                              | 47        |
| 2.5 Summary and Conclusions . . . . .                                  | 48        |
| <b>II Probability</b>  | <b>53</b> |

|   |           |
|---|-----------|
| <b>3 Fundamentals of Probability Theory</b>                 | <b>57</b> |
| 3.1 Building Blocks . . . . .                               | 58        |
| 3.2 Operations . . . . .                                    | 61        |
| 3.2.1 Events, Sets and Set Operations . . . . .             | 61        |
| 3.2.2 Set Functions . . . . .                               | 65        |
| 3.2.3 Probability Set Function . . . . .                    | 67        |
| 3.2.4 Final considerations . . . . .                        | 68        |
| 3.3 Probability . . . . .                                   | 69        |
| 3.3.1 The Calculus of Probability . . . . .                 | 69        |
| 3.3.2 Implications . . . . .                                | 71        |
| 3.4 Conditional Probability . . . . .                       | 72        |
| 3.4.1 Illustrating the Concept . . . . .                    | 72        |
| 3.4.2 Formalizing the Concept . . . . .                     | 73        |
| 3.4.3 Total Probability . . . . .                           | 74        |
| 3.4.4 Bayes' Rule . . . . .                                 | 76        |
| 3.5 Independence . . . . .                                  | 77        |
| 3.6 Summary and Conclusions . . . . .                       | 78        |
| <b>4 Random Variables and Distributions</b>                 | <b>89</b> |
| 4.1 Introduction and Definition . . . . .                   | 90        |
| 4.1.1 Mathematical Concept of the Random Variable . . . . . | 90        |
| 4.1.2 Practical Considerations . . . . .                    | 94        |
| 4.1.3 Types of Random Variables . . . . .                   | 94        |
| 4.2 Distributions . . . . .                                 | 95        |
| 4.2.1 Discrete Random Variables . . . . .                   | 95        |
| 4.2.2 Continuous Random Variables . . . . .                 | 98        |
| 4.2.3 The Probability Distribution Function . . . . .       | 100       |
| 4.3 Mathematical Expectation . . . . .                      | 102       |
| 4.3.1 Motivating the Definition . . . . .                   | 102       |
| 4.3.2 Definition and Properties . . . . .                   | 105       |
| 4.4 Characterizing Distributions . . . . .                  | 107       |
| 4.4.1 Moments of a Distributions . . . . .                  | 107       |
| 4.4.2 Moment Generating Function . . . . .                  | 113       |
| 4.4.3 Characteristic Function . . . . .                     | 115       |
| 4.4.4 Additional Distributional Characteristics . . . . .   | 116       |
| 4.4.5 Entropy . . . . .                                     | 119       |
| 4.4.6 Probability Bounds . . . . .                          | 119       |
| 4.5 Special Derived Probability Functions . . . . .         | 122       |
| 4.5.1 Survival Function . . . . .                           | 122       |
| 4.5.2 Hazard Function . . . . .                             | 123       |
| 4.5.3 Cumulative Hazard Function . . . . .                  | 124       |
| 4.6 Summary and Conclusions . . . . .                       | 124       |

|  |            |
|--|------------|
| <b>5 Multidimensional Random Variables</b>   | <b>137</b> |
| 5.1 Introduction and Definitions . . . . .   | 138        |
| 5.1.1 Perspectives . . . . .   | 138        |
| 5.1.2 2-Dimensional (Bivariate) Random Variables . . . . .                           | 139        |
| 5.1.3 Higher-Dimensional (Multivariate) Random Variables . . . . .                   | 140        |
| 5.2 Distributions of Several Random Variables . . . . .                              | 141        |
| 5.2.1 Joint Distributions . . . . .  | 141        |
| 5.2.2 Marginal Distributions . . . . .   | 144        |
| 5.2.3 Conditional Distributions . . . . .  | 147        |
| 5.2.4 General Extensions . . . . .   | 153        |
| 5.3 Distributional Characteristics of Jointly Distributed Random Variables . . . . . | 154        |
| 5.3.1 Expectations . . . . .   | 154        |
| 5.3.2 Covariance and Correlation . . . . .   | 157        |
| 5.3.3 Independence . . . . .   | 158        |
| 5.4 Summary and Conclusions . . . . .  | 163        |
| <b>6 Random Variable Transformations</b>   | <b>171</b> |
| 6.1 Introduction and Problem Definition . . . . .                                    | 172        |
| 6.2 Single Variable Transformations . . . . .  | 172        |
| 6.2.1 Discrete Case . . . . .  | 173        |
| 6.2.2 Continuous Case . . . . .  | 175        |
| 6.2.3 General Continuous Case . . . . .  | 176        |
| 6.2.4 Random Variable Sums . . . . .   | 177        |
| 6.3 Bivariate Transformations . . . . .  | 182        |
| 6.4 General Multivariate Transformations . . . . .                                   | 184        |
| 6.4.1 Square Transformations . . . . .   | 184        |
| 6.4.2 Non-Square Transformations . . . . .   | 185        |
| 6.4.3 Non-Monotone Transformations . . . . .   | 188        |
| 6.5 Summary and Conclusions . . . . .  | 188        |
| <b>7 Application Case Studies I: Probability</b>                                     | <b>197</b> |
| 7.1 Introduction . . . . .   | 198        |
| 7.2 Mendel and Heredity . . . . .  | 199        |
| 7.2.1 Background and Problem Definition . . . . .                                    | 199        |
| 7.2.2 Single Trait Experiments and Results . . . . .                                 | 201        |
| 7.2.3 Single trait analysis . . . . .  | 201        |
| 7.2.4 Multiple Traits and Independence . . . . .                                     | 205        |
| 7.2.5 Subsequent Experiments and Conclusions . . . . .                               | 208        |
| 7.3 World War II Warship Tactical Response Under Attack . . . . .                    | 209        |
| 7.3.1 Background and Problem Definition . . . . .                                    | 209        |
| 7.3.2 Approach and Results . . . . .   | 210        |
| 7.3.3 Final Comments . . . . .   | 212        |
| 7.4 Summary and Conclusions . . . . .  | 212        |

|  |            |
|--|------------|
| <b>III Distributions</b>   | <b>213</b> |
| <b>8 Ideal Models of Discrete Random Variables</b>               | <b>217</b> |
| 8.1 Introduction . . . . .                                       | 218        |
| 8.2 The Discrete Uniform Random Variable . . . . .               | 219        |
| 8.2.1 Basic Characteristics and Model . . . . .                  | 219        |
| 8.2.2 Applications . . . . .                                     | 220        |
| 8.3 The Bernoulli Random Variable . . . . .                      | 221        |
| 8.3.1 Basic Characteristics . . . . .                            | 221        |
| 8.3.2 Model Development . . . . .                                | 221        |
| 8.3.3 Important Mathematical Characteristics . . . . .           | 222        |
| 8.4 The Hypergeometric Random Variable . . . . .                 | 222        |
| 8.4.1 Basic Characteristics . . . . .                            | 222        |
| 8.4.2 Model Development . . . . .                                | 223        |
| 8.4.3 Important Mathematical Characteristics . . . . .           | 224        |
| 8.4.4 Applications . . . . .                                     | 224        |
| 8.5 The Binomial Random Variable . . . . .                       | 225        |
| 8.5.1 Basic Characteristics . . . . .                            | 225        |
| 8.5.2 Model Development . . . . .                                | 225        |
| 8.5.3 Important Mathematical Characteristics . . . . .           | 226        |
| 8.5.4 Applications . . . . .                                     | 227        |
| 8.6 Extensions and Special Cases of the Binomial Random Variable | 230        |
| 8.6.1 Trinomial Random Variable . . . . .                        | 230        |
| 8.6.2 Multinomial Random Variable . . . . .                      | 231        |
| 8.6.3 Negative Binomial Random Variable . . . . .                | 232        |
| 8.6.4 Geometric Random Variable . . . . .                        | 234        |
| 8.7 The Poisson Random Variable . . . . .                        | 236        |
| 8.7.1 The Limiting Form of a Binomial Random Variable .          | 236        |
| 8.7.2 First Principles Derivation . . . . .                      | 237        |
| 8.7.3 Important Mathematical Characteristics . . . . .           | 239        |
| 8.7.4 Applications . . . . .                                     | 240        |
| 8.8 Summary and Conclusions . . . . .                            | 243        |
| <b>9 Ideal Models of Continuous Random Variables</b>             | <b>257</b> |
| 9.1 Gamma Family Random Variables . . . . .                      | 259        |
| 9.1.1 The Exponential Random Variable . . . . .                  | 260        |
| 9.1.2 The Gamma Random Variable . . . . .                        | 264        |
| 9.1.3 The Chi-Square Random Variable . . . . .                   | 271        |
| 9.1.4 The Weibull Random Variable . . . . .                      | 272        |
| 9.1.5 The Generalized Gamma Model . . . . .                      | 276        |
| 9.1.6 The Poisson-Gamma Mixture Distribution . . . . .           | 276        |
| 9.2 Gaussian Family Random Variables . . . . .                   | 278        |
| 9.2.1 The Gaussian (Normal) Random Variable . . . . .            | 279        |
| 9.2.2 The Standard Normal Random Variable . . . . .              | 290        |
| 9.2.3 The Lognormal Random Variable . . . . .                    | 292        |

|           |  |            |
|-----------|--|------------|
| 9.2.4     | The Rayleigh Random Variable . . . . .                               | 297        |
| 9.2.5     | The Generalized Gaussian Model . . . . .                             | 300        |
| 9.3       | Ratio Family Random Variables . . . . .                              | 300        |
| 9.3.1     | The Beta Random Variable . . . . .                                   | 301        |
| 9.3.2     | Extensions and Special Cases of the Beta Random Variable . . . . .   | 307        |
| 9.3.3     | The (Continuous) Uniform Random Variable . . . . .                   | 308        |
| 9.3.4     | Fisher's F Random Variable . . . . .                                 | 309        |
| 9.3.5     | Student's t Random Variable . . . . .                                | 311        |
| 9.3.6     | The Cauchy Random Variable . . . . .                                 | 314        |
| 9.4       | Summary and Conclusions . . . . .                                    | 316        |
| <b>10</b> | <b>Information, Entropy and Probability Models</b>                   | <b>335</b> |
| 10.1      | Uncertainty and Information . . . . .                                | 336        |
| 10.1.1    | Basic Concepts . . . . .   | 336        |
| 10.1.2    | Quantifying Information . . . . .                                    | 337        |
| 10.2      | Entropy . . . . .  | 338        |
| 10.2.1    | Discrete Random Variables . . . . .                                  | 338        |
| 10.2.2    | Continuous Random Variables . . . . .                                | 340        |
| 10.3      | Maximum Entropy Principles for Probability Modeling . . . . .        | 344        |
| 10.4      | Some Maximum Entropy Models . . . . .                                | 344        |
| 10.4.1    | Discrete Random Variable; Known Range . . . . .                      | 345        |
| 10.4.2    | Discrete Random Variable; Known Mean . . . . .                       | 346        |
| 10.4.3    | Continuous Random Variable; Known Range . . . . .                    | 348        |
| 10.4.4    | Continuous Random Variable; Known Mean . . . . .                     | 349        |
| 10.4.5    | Continuous Random Variable; Known Mean and Variance . . . . .        | 350        |
| 10.4.6    | Continuous Random Variable; Known Range, Mean and Variance . . . . . | 351        |
| 10.5      | Maximum Entropy Models from General Expectations . . . . .           | 351        |
| 10.5.1    | Single Expectations . . . . .  | 351        |
| 10.5.2    | Multiple Expectations . . . . .                                      | 353        |
| 10.6      | Summary and Conclusions . . . . .                                    | 354        |
| <b>11</b> | <b>Application Case Studies II: In-Vitro Fertilization</b>           | <b>363</b> |
| 11.1      | Introduction . . . . .   | 364        |
| 11.2      | In-Vitro Fertilization and Multiple Births . . . . .                 | 365        |
| 11.2.1    | Background and Problem Definition . . . . .                          | 365        |
| 11.2.2    | Clinical Studies and Recommended Guidelines . . . . .                | 367        |
| 11.3      | Probability Modeling and Analysis . . . . .                          | 371        |
| 11.3.1    | Model Postulate . . . . .  | 371        |
| 11.3.2    | Prediction . . . . .   | 372        |
| 11.3.3    | Estimation . . . . .   | 373        |
| 11.4      | Binomial Model Validation . . . . .                                  | 375        |
| 11.4.1    | Overview and Study Characteristics . . . . .                         | 375        |

|  |            |
|--|------------|
| 11.4.2 Binomial Model versus Clinical Data . . . . .                                   | 377        |
| 11.5 Problem Solution: Model-based IVF Optimization and Analysis                       | 384        |
| 11.5.1 Optimization . . . . .  | 385        |
| 11.5.2 Model-based Analysis . . . . .  | 386        |
| 11.5.3 Patient Categorization and Theoretical Analysis of Treatment Outcomes . . . . . | 390        |
| 11.6 Sensitivity Analysis . . . . .  | 392        |
| 11.6.1 General Discussion . . . . .  | 392        |
| 11.6.2 Theoretical Sensitivity Analysis . . . . .                                      | 394        |
| 11.7 Summary and Conclusions . . . . .   | 395        |
| 11.7.1 Final Wrap-up . . . . .   | 395        |
| 11.7.2 Conclusions and Perspectives on Previous Studies and Guidelines . . . . .       | 397        |
| <b>IV Statistics</b>   | <b>403</b> |
| <b>12 Introduction to Statistics</b>   | <b>407</b> |
| 12.1 From Probability to Statistics . . . . .  | 408        |
| 12.1.1 Random Phenomena and Finite Data Sets . . . . .                                 | 408        |
| 12.1.2 Finite Data Sets and Statistical Analysis . . . . .                             | 411        |
| 12.1.3 Probability, Statistics and Design of Experiments . . . . .                     | 414        |
| 12.1.4 Statistical Analysis . . . . .  | 415        |
| 12.2 Variable and Data Types . . . . .   | 417        |
| 12.3 Graphical Methods of Descriptive Statistics . . . . .                             | 419        |
| 12.3.1 Bar Charts and Pie Charts . . . . .   | 419        |
| 12.3.2 Frequency Distributions . . . . .   | 424        |
| 12.3.3 Box Plots . . . . .   | 427        |
| 12.3.4 Scatter Plots . . . . .   | 431        |
| 12.4 Numerical Descriptions . . . . .  | 436        |
| 12.4.1 Theoretical Measures of Central Tendency . . . . .                              | 436        |
| 12.4.2 Measures of Central Tendency: Sample Equivalents . . . . .                      | 438        |
| 12.4.3 Measures of Variability . . . . .   | 440        |
| 12.4.4 Supplementing Numerics with Graphics . . . . .                                  | 442        |
| 12.5 Summary and Conclusions . . . . .   | 446        |
| <b>13 Sampling</b>   | <b>459</b> |
| 13.1 Introductory Concepts . . . . .   | 460        |
| 13.1.1 The Random Sample . . . . .   | 460        |
| 13.1.2 The “Statistic” and its Distribution . . . . .                                  | 461        |
| 13.2 The Distribution of Functions of Random Variables . . . . .                       | 463        |
| 13.2.1 General Overview . . . . .  | 463        |
| 13.2.2 Some Important Sampling Distribution Results . . . . .                          | 463        |
| 13.3 Sampling Distribution of The Mean . . . . .                                       | 465        |
| 13.3.1 Underlying Probability Distribution Known . . . . .                             | 465        |
| 13.3.2 Underlying Probability Distribution Unknown . . . . .                           | 467        |

|  |            |
|--|------------|
| 13.3.3 Limiting Distribution of the Mean . . . . .                             | 467        |
| 13.3.4 $\sigma$ Unknown . . . . .  | 470        |
| 13.4 Sampling Distribution of the Variance . . . . .                           | 473        |
| 13.5 Summary and Conclusions . . . . .   | 476        |
| <b>14 Estimation</b>   | <b>487</b> |
| 14.1 Introductory Concepts . . . . .   | 488        |
| 14.1.1 An Illustration . . . . .   | 488        |
| 14.1.2 Problem Definition and Key Concepts . . . . .                           | 489        |
| 14.2 Criteria for Selecting Estimators . . . . .                               | 490        |
| 14.2.1 Unbiasedness . . . . .  | 490        |
| 14.2.2 Efficiency . . . . .  | 491        |
| 14.2.3 Consistency . . . . .   | 492        |
| 14.3 Point Estimation Methods . . . . .  | 493        |
| 14.3.1 Method of Moments . . . . .   | 493        |
| 14.3.2 Maximum Likelihood . . . . .  | 496        |
| 14.4 Precision of Point Estimates . . . . .                                    | 503        |
| 14.5 Interval Estimates . . . . .  | 506        |
| 14.5.1 General Principles . . . . .  | 506        |
| 14.5.2 Mean of a Normal Population; $\sigma$ Known . . . . .                   | 507        |
| 14.5.3 Mean of a Normal Population; $\sigma$ Unknown . . . . .                 | 508        |
| 14.5.4 Variance of a Normal Population . . . . .                               | 510        |
| 14.5.5 Difference of Two Normal Populations Means . . . . .                    | 512        |
| 14.5.6 Interval Estimates for Parameters from other Popula-<br>tions . . . . . | 514        |
| 14.6 Bayesian Estimation . . . . .   | 518        |
| 14.6.1 Background . . . . .  | 518        |
| 14.6.2 Basic Concept . . . . .   | 519        |
| 14.6.3 Bayesian Estimation Results . . . . .                                   | 520        |
| 14.6.4 A Simple Illustration . . . . .   | 521        |
| 14.6.5 Discussion . . . . .  | 524        |
| 14.7 Summary and Conclusions . . . . .   | 527        |
| <b>15 Hypothesis Testing</b>   | <b>551</b> |
| 15.1 Introduction . . . . .  | 552        |
| 15.2 Basic Concepts . . . . .  | 554        |
| 15.2.1 Terminology and Definitions . . . . .                                   | 554        |
| 15.2.2 General Procedure . . . . .   | 560        |
| 15.3 Concerning Single Mean of a Normal Population . . . . .                   | 561        |
| 15.3.1 $\sigma$ Known; the “z-test” . . . . .                                  | 563        |
| 15.3.2 $\sigma$ Unknown; the “t-test” . . . . .                                | 570        |
| 15.3.3 Confidence Intervals and Hypothesis Tests . . . . .                     | 575        |
| 15.4 Concerning Two Normal Population Means . . . . .                          | 576        |
| 15.4.1 Population Standard Deviations Known . . . . .                          | 576        |
| 15.4.2 Population Standard Deviations Unknown . . . . .                        | 578        |

|   |            |
|---|------------|
| 15.4.3 Paired Differences . . . . .                             | 585        |
| 15.5 Determining $\beta$ , Power, and Sample Size . . . . .     | 591        |
| 15.5.1 $\beta$ and Power . . . . .                              | 591        |
| 15.5.2 Sample Size . . . . .                                    | 593        |
| 15.5.3 $\beta$ and Power for Lower-Tailed and Two-Sided Tests . | 598        |
| 15.5.4 General Power and Sample Size Considerations . . . . .   | 599        |
| 15.6 Concerning Variances of Normal Populations . . . . .       | 600        |
| 15.6.1 Single Variance . . . . .                                | 601        |
| 15.6.2 Two Variances . . . . .                                  | 603        |
| 15.7 Concerning Proportions . . . . .                           | 606        |
| 15.7.1 Single Population Proportion . . . . .                   | 607        |
| 15.7.2 Two Population Proportions . . . . .                     | 610        |
| 15.8 Concerning Non-Gaussian Populations . . . . .              | 613        |
| 15.8.1 Large Sample Test for Means . . . . .                    | 613        |
| 15.8.2 Small Sample Tests . . . . .                             | 614        |
| 15.9 Likelihood Ratio Tests . . . . .                           | 616        |
| 15.9.1 General Principles . . . . .                             | 616        |
| 15.9.2 Special Cases . . . . .                                  | 619        |
| 15.9.3 Asymptotic Distribution for $\Lambda$ . . . . .          | 622        |
| 15.10 Discussion . . . . .                                      | 623        |
| 15.11 Summary and Conclusions . . . . .                         | 624        |
| <b>16 Regression Analysis</b>                                   | <b>647</b> |
| 16.1 Introductory Concepts . . . . .                            | 648        |
| 16.1.1 Dependent and Independent Variables . . . . .            | 650        |
| 16.1.2 The Principle of Least Squares . . . . .                 | 651        |
| 16.2 Simple Linear Regression . . . . .                         | 652        |
| 16.2.1 One-Parameter Model . . . . .                            | 652        |
| 16.2.2 Two-Parameter Model . . . . .                            | 653        |
| 16.2.3 Properties of OLS Estimators . . . . .                   | 660        |
| 16.2.4 Confidence Intervals . . . . .                           | 661        |
| 16.2.5 Hypothesis Testing . . . . .                             | 664        |
| 16.2.6 Prediction and Prediction Intervals . . . . .            | 668        |
| 16.2.7 Coefficient of Determination and the F-Test . . . . .    | 671        |
| 16.2.8 Relation to the Correlation Coefficient . . . . .        | 676        |
| 16.2.9 Mean-Centered Model . . . . .                            | 677        |
| 16.2.10 Residual Analysis . . . . .                             | 678        |
| 16.3 “Intrinsically” Linear Regression . . . . .                | 682        |
| 16.3.1 Linearity in Regression Models . . . . .                 | 682        |
| 16.3.2 Variable Transformations . . . . .                       | 685        |
| 16.4 Multiple Linear Regression . . . . .                       | 686        |
| 16.4.1 General Least Squares . . . . .                          | 687        |
| 16.4.2 Matrix Methods . . . . .                                 | 688        |
| 16.4.3 Some Important Special Cases . . . . .                   | 694        |
| 16.4.4 Recursive Least Squares . . . . .                        | 698        |

|           |  |            |
|-----------|--|------------|
| 16.5      | Polynomial Regression . . . . .                        | 700        |
| 16.5.1    | General Considerations . . . . .                       | 700        |
| 16.5.2    | Orthogonal Polynomial Regression . . . . .             | 704        |
| 16.6      | Summary and Conclusions . . . . .                      | 710        |
| <b>17</b> | <b>Probability Model Validation</b>                    | <b>731</b> |
| 17.1      | Introduction . . . . .                                 | 732        |
| 17.2      | Probability Plots . . . . .                            | 733        |
| 17.2.1    | Basic Principles . . . . .                             | 733        |
| 17.2.2    | Transformations and Specialized Graph Papers . . . . . | 734        |
| 17.2.3    | Modern Probability Plots . . . . .                     | 736        |
| 17.2.4    | Applications . . . . .                                 | 737        |
| 17.3      | Chi-Squared Goodness-of-fit Test . . . . .             | 739        |
| 17.3.1    | Basic Principles . . . . .                             | 739        |
| 17.3.2    | Properties and Application . . . . .                   | 742        |
| 17.4      | Summary and Conclusions . . . . .                      | 745        |
| <b>18</b> | <b>Nonparametric Methods</b>                           | <b>757</b> |
| 18.1      | Introduction . . . . .                                 | 758        |
| 18.2      | Single Population . . . . .                            | 760        |
| 18.2.1    | One-Sample Sign Test . . . . .                         | 760        |
| 18.2.2    | One-Sample Wilcoxon Signed Rank Test . . . . .         | 763        |
| 18.3      | Two Populations . . . . .                              | 765        |
| 18.3.1    | Two-Sample Paired Test . . . . .                       | 766        |
| 18.3.2    | Mann-Whitney-Wilcoxon Test . . . . .                   | 766        |
| 18.4      | Probability Model Validation . . . . .                 | 770        |
| 18.4.1    | The Kolmogorov-Smirnov Test . . . . .                  | 770        |
| 18.4.2    | The Anderson-Darling Test . . . . .                    | 771        |
| 18.5      | A Comprehensive Illustrative Example . . . . .         | 772        |
| 18.5.1    | Probability Model Postulate and Validation . . . . .   | 772        |
| 18.5.2    | Mann-Whitney-Wilcoxon Test . . . . .                   | 775        |
| 18.6      | Summary and Conclusions . . . . .                      | 777        |
| <b>19</b> | <b>Design of Experiments</b>                           | <b>791</b> |
| 19.1      | Introductory Concepts . . . . .                        | 793        |
| 19.1.1    | Experimental Studies and Design . . . . .              | 793        |
| 19.1.2    | Phases of Efficient Experimental Studies . . . . .     | 794        |
| 19.1.3    | Problem Definition and Terminology . . . . .           | 795        |
| 19.2      | Analysis of Variance . . . . .                         | 796        |
| 19.3      | Single Factor Experiments . . . . .                    | 797        |
| 19.3.1    | One-Way Classification . . . . .                       | 797        |
| 19.3.2    | Kruskal-Wallis Nonparametric Test . . . . .            | 805        |
| 19.3.3    | Two-Way Classification . . . . .                       | 805        |
| 19.3.4    | Other Extensions . . . . .                             | 811        |
| 19.4      | Two-Factor Experiments . . . . .                       | 811        |

|           |  |            |
|-----------|--|------------|
| 19.5      | General Multi-factor Experiments . . . . .                             | 814        |
| 19.6      | $2^k$ Factorial Experiments and Design . . . . .                       | 814        |
| 19.6.1    | Overview . . . . .   | 814        |
| 19.6.2    | Design and Analysis . . . . .  | 816        |
| 19.6.3    | Procedure . . . . .  | 817        |
| 19.6.4    | Closing Remarks . . . . .  | 821        |
| 19.7      | Screening Designs: Fractional Factorial . . . . .                      | 822        |
| 19.7.1    | Rationale . . . . .  | 822        |
| 19.7.2    | Illustrating the Mechanics . . . . .                                   | 822        |
| 19.7.3    | General characteristics . . . . .                                      | 823        |
| 19.7.4    | Design and Analysis . . . . .  | 825        |
| 19.7.5    | A Practical Illustrative Example . . . . .                             | 827        |
| 19.8      | Screening Designs: Plackett-Burman . . . . .                           | 832        |
| 19.8.1    | Primary Characteristics . . . . .                                      | 833        |
| 19.8.2    | Design and Analysis . . . . .  | 833        |
| 19.9      | Response Surface Designs . . . . .                                     | 834        |
| 19.9.1    | Characteristics . . . . .  | 834        |
| 19.9.2    | Response Surface Designs . . . . .                                     | 835        |
| 19.9.3    | Design and Analysis . . . . .  | 836        |
| 19.10     | Introduction to Optimal Designs . . . . .                              | 837        |
| 19.10.1   | Background . . . . .   | 837        |
| 19.10.2   | “Alphabetic” Optimal Designs . . . . .                                 | 838        |
| 19.11     | Summary and Conclusions . . . . .                                      | 839        |
| <b>20</b> | <b>Application Case Studies III: Statistics</b>                        | <b>855</b> |
| 20.1      | Introduction . . . . .   | 856        |
| 20.2      | Prussian Army Death-by-Horse kicks . . . . .                           | 857        |
| 20.2.1    | Background and Data . . . . .  | 857        |
| 20.2.2    | Parameter Estimation and Model Validation . . . . .                    | 859        |
| 20.2.3    | Recursive Bayesian Estimation . . . . .                                | 860        |
| 20.3      | WW II Aerial Bombardment of London . . . . .                           | 868        |
| 20.4      | US Population Dynamics: 1790-2000 . . . . .                            | 870        |
| 20.4.1    | Background and Data . . . . .  | 870        |
| 20.4.2    | “Truncated Data” Modeling and Evaluation . . . . .                     | 872        |
| 20.4.3    | Full Data Set Modeling and Evaluation . . . . .                        | 873        |
| 20.4.4    | Hypothesis Testing Concerning Average Population Growth Rate . . . . . | 876        |
| 20.5      | Process Optimization . . . . .   | 879        |
| 20.5.1    | Problem Definition and Background . . . . .                            | 879        |
| 20.5.2    | Experimental Strategy and Results . . . . .                            | 879        |
| 20.5.3    | Analysis . . . . .   | 880        |
| 20.6      | Summary and Conclusions . . . . .                                      | 889        |
| <b>V</b>  | <b>Applications</b>  | <b>895</b> |

|  |            |
|--|------------|
| <b>21 Reliability and Life Testing</b>                                   | <b>899</b> |
| 21.1 Introduction . . . . .  | 900        |
| 21.2 System Reliability . . . . .  | 901        |
| 21.2.1 Simple Systems . . . . .  | 901        |
| 21.2.2 Complex Systems . . . . .   | 906        |
| 21.3 System Lifetime and Failure-Time Distributions . . . . .            | 911        |
| 21.3.1 Characterizing Time-to-Failure . . . . .                          | 911        |
| 21.3.2 Probability Models for Distribution of Failure Times .            | 913        |
| 21.4 The Exponential Reliability Model . . . . .                         | 914        |
| 21.4.1 Component Characteristics . . . . .                               | 914        |
| 21.4.2 Series Configuration . . . . .                                    | 915        |
| 21.4.3 Parallel Configuration . . . . .                                  | 916        |
| 21.4.4 $m$ -of- $n$ Parallel Systems . . . . .                           | 917        |
| 21.5 The Weibull Reliability Model . . . . .                             | 918        |
| 21.6 Life Testing . . . . .  | 919        |
| 21.6.1 The Exponential Model . . . . .                                   | 919        |
| 21.6.2 The Weibull Model . . . . .                                       | 922        |
| 21.7 Summary and Conclusions . . . . .                                   | 923        |
| <b>22 Quality Assurance and Control</b>                                  | <b>933</b> |
| 22.1 Introduction . . . . .  | 934        |
| 22.2 Acceptance Sampling . . . . .                                       | 936        |
| 22.2.1 Basic Principles . . . . .  | 936        |
| 22.2.2 Determining a Sampling Plan . . . . .                             | 938        |
| 22.3 Process and Quality Control . . . . .                               | 944        |
| 22.3.1 Underlying Philosophy . . . . .                                   | 944        |
| 22.3.2 Statistical Process Control . . . . .                             | 944        |
| 22.3.3 Basic Control Charts . . . . .                                    | 946        |
| 22.3.4 Enhancements . . . . .  | 958        |
| 22.4 Chemical Process Control . . . . .                                  | 964        |
| 22.4.1 Preliminary Considerations . . . . .                              | 964        |
| 22.4.2 Statistical Process Control (SPC) Perspective . . . . .           | 965        |
| 22.4.3 Engineering/Automatic Process Control (APC) Perspective . . . . . | 966        |
| 22.4.4 SPC or APC . . . . .  | 967        |
| 22.5 Process and Parameter Design . . . . .                              | 969        |
| 22.5.1 Basic Principles . . . . .  | 969        |
| 22.5.2 A Theoretical Rationale . . . . .                                 | 970        |
| 22.6 Summary and Conclusions . . . . .                                   | 971        |
| <b>23 Introduction to Multivariate Analysis</b>                          | <b>977</b> |
| 23.1 Multivariate Probability Models . . . . .                           | 978        |
| 23.1.1 Introduction . . . . .  | 978        |
| 23.1.2 The Multivariate Normal Distribution . . . . .                    | 979        |
| 23.1.3 The Wishart Distribution . . . . .                                | 981        |

|  |             |
|--|-------------|
| 23.1.4 Hotelling's $T$ -Squared Distribution . . . . . | 982         |
| 23.1.5 The Wilks Lambda Distribution . . . . .         | 982         |
| 23.1.6 The Dirichlet Distribution . . . . .            | 983         |
| 23.2 Multivariate Data Analysis . . . . .              | 984         |
| 23.3 Principal Components Analysis . . . . .           | 985         |
| 23.3.1 Basic Principles of PCA . . . . .               | 986         |
| 23.3.2 Main Characteristics of PCA . . . . .           | 990         |
| 23.3.3 Illustrative example . . . . .                  | 991         |
| 23.3.4 Other Applications of PCA . . . . .             | 999         |
| 23.4 Summary and Conclusions . . . . .                 | 1002        |
| <b>Appendix</b>  | <b>1005</b> |
| <b>Index</b>   | <b>1009</b> |

# Chapter 0

---

## Prelude

|     |                               |   |
|-----|-------------------------------|---|
| 0.1 | Approach Philosophy .....     | 1 |
| 0.2 | Four basic principles .....   | 3 |
| 0.3 | Summary and Conclusions ..... | 5 |

*Rem tene; verba sequentur.  
(Grasp the subject; the words will follow.)*

Cato the Elder (234–149 BC)

From weather forecasts and life insurance premiums for non-smokers to clinical tests of experimental drugs and defect rates in manufacturing facilities, and in numerous other ways, *randomly varying phenomena* exert a subtle but pervasive influence on everyday life. In most cases, one can be blissfully ignorant of the true implications of the presence of such phenomena without consequence. In science and engineering, however, the influence of randomly varying phenomena can be such that even apparently simple problems can become dramatically complicated by the presence of random variability—demanding special methods and analysis tools for obtaining valid and useful solutions.

*The primary aim of this book is to provide the reader with the basic fundamental principles, methods, and tools for formulating and solving engineering problems involving randomly varying phenomena.*

Since this aim can be achieved in several different ways, this chapter is devoted to presenting this book’s approach philosophy.

---

### 0.1 Approach Philosophy

Engineers are typically well-trained in the “art” of problem formulation and problem solving when *all* the entities involved are considered *deterministic* in character. However, many problems of practical importance involve

*randomly varying phenomena* of one sort or another; and the vast majority of such problems cannot always be idealized and reduced to the more familiar “deterministic” types without destroying the very essence of the problem. For example, in determining which of two catalysts A or B provides the greater yield in a chemical manufacturing process, it is well-known that the respective yields  $Y_A$  and  $Y_B$ , *as observed experimentally*, are randomly varying quantities. Chapter 1 presents a full-scale discussion of this problem. For now, we simply note that with catalyst A, fifty different experiments performed under essentially identical conditions will result in fifty different values (*realizations*) for  $Y_A$ . Similarly for catalyst B, one obtains fifty distinct values for  $Y_B$  from fifty different experiments replicated under identical conditions. The first 10 experimental data points for this example are shown in the table below.

| $Y_A$ % | $Y_B$ % |
|---------|---------|
| 74.04   | 75.75   |
| 75.29   | 68.41   |
| 75.62   | 74.19   |
| 75.91   | 68.10   |
| 77.21   | 68.10   |
| 75.07   | 69.23   |
| 74.23   | 70.14   |
| 74.92   | 69.22   |
| 76.57   | 74.17   |
| 77.77   | 70.23   |

Observe that because of the variability inherent in the data, some of the  $Y_A$  values are greater than some of the  $Y_B$  values; but the converse is also true—some  $Y_B$  values are greater than some  $Y_A$  values. So how does one determine—reliably and confidently—which catalyst (if any) really provides the greater yield? Clearly, special methods and analysis tools are required for handling this apparently simple problem: the deterministic idealization of comparing a *single* observed value of  $Y_A$  (say the first entry, 74.04) with a corresponding *single* observed value of  $Y_B$  (in this case 75.75) is incapable of producing a valid answer. The primary essence of this problem is the variability inherent in the data which masks the fact that one catalyst does in fact provide the greater yield.

This book takes a more fundamental, “first-principles” approach to the issue of dealing with random variability and uncertainty in engineering problems. This is in contrast to the typical “engineering statistics” approach on the one hand, or the “problem-specific” approach on the other. With the former approach, most of the emphasis is on *how* to use certain popular statistical techniques to solve *some* of the most commonly encountered engineering problems, with little or no discussion of *why* the techniques are effective. With the latter approach, a particular topic (say “Design of Experiments”) is selected and dealt with in depth, and the appropriate statistical tools are presented and discussed within the context of the specific problem at the core of the

selected topic. By definition, such an approach excludes all other topics that may be of practical interest, opting to make up in depth what it gives up in breadth.

*The approach taken in this book is based on the premise that emphasizing fundamentals and basic principles, and then illustrating these with examples, equips the reader with the means of dealing with a range of problems wider than that explicitly covered in the book.*

This approach philosophy is based on the four basic principles discussed next.

---

## 0.2 Four basic principles

1. *If characterized properly, random phenomena are subject to rigorous mathematical analysis in much the same manner as deterministic phenomena.*

Random phenomena are so-called because they show no apparent regularity, appearing to occur haphazardly—totally at random; the observed variations do not seem to obey any discernible rational laws and therefore appear to be entirely unpredictable. However, the unpredictable irregularities of the individual observations (or, more generally, the “detail”) of random phenomena in fact co-exist with surprisingly predictable ensemble, or aggregate, behavior. This fact makes rigorous analysis possible; it also provides the basis for employing the concept and calculus of “probability” to develop a systematic framework for characterizing random phenomena in terms of “probability distribution functions.”

The first order of business is therefore to seek to understand random phenomena and to develop techniques for characterizing them appropriately. Part I, titled *FOUNDATIONS: Understanding Random Variability*, and Part II, titled *PROBABILITY: Characterizing Random Variability*, are devoted to these respective tasks. Ultimately, probability—and the probability distribution function—are introduced as the theoretical constructs for efficiently describing *our knowledge* of the real-world phenomena in question.

2. *By focusing on the underlying phenomenological mechanisms , it is possible to develop appropriate theoretical characterizations of random phenomena in terms of ideal models of the observed variability.*

Within the probabilistic framework, the ensemble, or aggregate behavior

of the random phenomenon in question is characterized by its “probability distribution function.” In much the same way that theoretical mathematical models are derived from “first-principles” for deterministic phenomena, it is also possible to derive these theoretical probability distribution functions as ideal models that describe our knowledge of the underlying random phenomena. Part III, titled *DISTRIBUTIONS: Modeling Random Variability*, is devoted to the important tasks of developing and analyzing ideal probability models for many random phenomena of practical interest. The end result is a collection of probability distribution functions each derived directly from—and hence explicitly linked to—the underlying random phenomenological mechanisms.

*3. The ensemble (or aggregate) characterization provided by ideal probability models can be used successfully to develop the theoretical basis for solving real problems where one is always limited to dealing with an incomplete collection of individual observations—never the entire aggregate.*

A key defining characteristic of random phenomena is that specific outcomes or observations cannot be predicted with *absolute certainty*. With probabilistic analysis, this otherwise “impossible task” of predicting the unpredictable individual observation or outcome is simply replaced by the analytical task of determining the mathematical *probability* of its occurrence. In many practical problems involving random phenomena, however, there is no avoiding this “impossible task;” one is required to deal with, and make decisions about, individual observations, and must therefore confront the inevitable uncertainty that will always be associated with such decisions. *Statistical Theory*, using the aggregate descriptions of probability theory, provides a rational basis not only for making these predictions and decisions about individual observations with confidence, but also for quantifying the degree of uncertainty associated with such decisions.

Part IV, titled *STATISTICS: Quantifying Random Variability*, is devoted to elucidating statistical principles and concepts required for dealing effectively with data as collections of individual observations from random phenomena.

*4. The usefulness and broad applicability of the fundamental principles, analysis methods, and tools provided by probability and statistics are best illustrated with several actual example topics of engineering applications involving random phenomena.*

The manifestations of random phenomena in problems of practical interest are countless, and the range of such problems is itself quite broad: from simple data analysis and experimental designs, to polynomial curve-fitting, and empirical modeling of complex dynamic systems; from quality assurance and control, to state and parameter estimation, and process monitoring and diagnosis, . . . etc. The topical headings under which such problems may be organized—Design of Experiments; Regression Analysis; Time Series Analysis; etc—are numerous, and many books have been devoted to each one of

them. Clearly then, the sheer vastness of the subject matter of engineering applications of probability and statistics renders completely unreasonable any hope of comprehensive coverage in a single introductory text.

Nevertheless, how probability and statistics are employed in practice to deal successfully with various problems created by random variability and uncertainty can be discussed in such a way as to equip the student with the tools needed to approach, with confidence, other problems that are not addressed explicitly in this book.

Part V, titled *APPLICATIONS: Dealing with Random Variability in Practice*, consists of three chapters each devoted to a specific application topic of importance in engineering practice. Entire books have been written, and entire courses taught, on each of the topics to which we will devote only one chapter; the coverage is therefore designed to be more illustrative than comprehensive, providing the basis for absorbing and employing more efficiently, the more extensive material presented in these other books or courses.

---

### 0.3 Summary and Conclusions

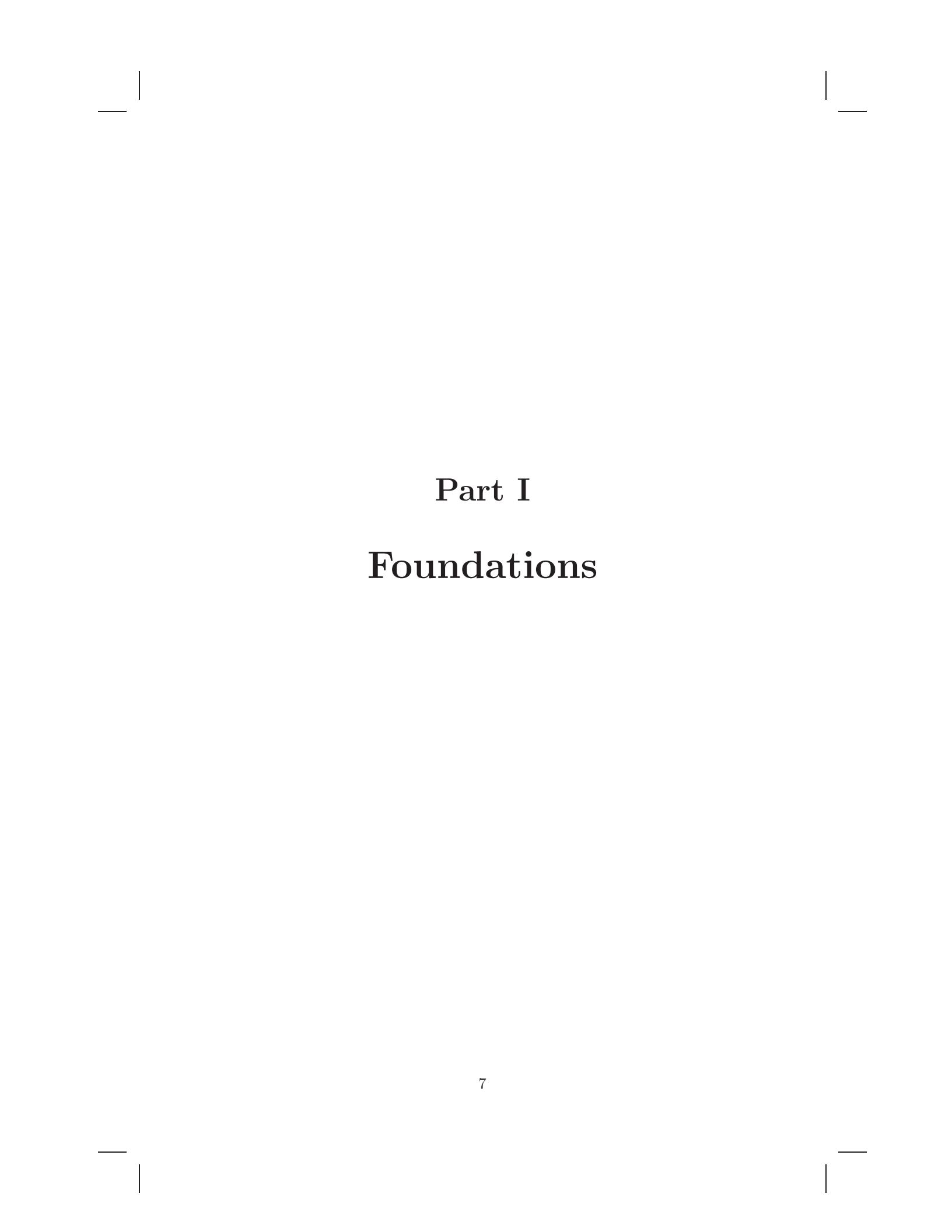
This chapter has been primarily concerned with setting forth this book's approach to presenting the fundamentals and engineering applications of probability and statistics. The four basic principles on which the more fundamental, "first principles" approach is based were presented, providing the rationale for the scope and organization of the material to be presented in the rest of the book.

The approach is designed to produce the following result:

*A course of study based on this book should provide the reader with a reasonable fundamental understanding of random phenomena, a working knowledge of how to model and analyze such phenomena, and facility with using probability and statistics to cope with random variability and uncertainty in some key engineering problems.*

The book should also prepare the student to absorb and employ the material presented in more problem-specific courses such as Design of Experiments, Time Series Analysis, Regression Analysis, Statistical Process Control, etc, a bit more efficiently.





# Part I

# Foundations

—

|

—

|

—

|

—

|

---

## Part I: Foundations

*Understanding Random Variability*

---

*I shall light a candle of understanding in thine heart which shall not be put out.*

*Apocrypha: I Esdras 14:25*

# Part I: Foundations

*Understanding Random Variability*

- **Chapter 1:** Two Motivating Examples
- **Chapter 2:** Random Phenomena, Variability and Uncertainty

# Chapter 1

---

## Two Motivating Examples

|                                |  |    |
|--------------------------------|--|----|
| 1.1                            | Yield Improvement in a Chemical Process .....                  | 11 |
| 1.1.1                          | The Problem .....  | 12 |
| Mathematical Formulation ..... | 12   |    |
| 1.1.2                          | The Essence of the Problem .....                               | 14 |
| 1.1.3                          | Preliminary “Intuitive” Notions .....                          | 14 |
| 1.2                            | Quality Assurance in a Glass Sheet Manufacturing Process ..... | 15 |
| 1.3                            | Outline of a Systematic Approach .....                         | 17 |
| 1.3.1                          | Group Classification and Frequency Distributions .....         | 18 |
| 1.3.2                          | Theoretical Distributions .....                                | 22 |
| A Preview .....                | 23   |    |
| 1.4                            | Summary and Conclusions .....                                  | 25 |
|                                | REVIEW QUESTIONS .....   | 26 |
|                                | EXERCISES .....  | 27 |
|                                | APPLICATION PROBLEMS .....                                     | 28 |

*And coming events cast their shadows before.  
(Lochiel’s warning.)*

Thomas Campbell (1777–1844)

When random variability is genuinely intrinsic to a problem, uncertainty becomes inevitable, but the problem can still be solved systematically and with *confidence*. This — the underlying theme of “Applied Probability and Statistics”— is what this chapter seeks to illustrate with two representative examples. The example problems and the accompanying discussions are intended to serve two main purposes: (i) illustrate the sort of complications caused by the presence of random components in practical problems; and (ii) demonstrate (qualitatively for now) how to solve such problems by formulating them properly and employing appropriate methods and tools. The primary value of this chapter is as a vehicle for placing in context this book’s approach to analyzing randomly varying phenomena in engineering and science. It allows us to preview and motivate the key concepts to be developed fully in the remaining chapters.

## 1.1 Yield Improvement in a Chemical Process

To an engineer or scientist, determining which of two numbers is larger, and by how much, is trivial, in principle requiring no more than the elementary arithmetic operation of subtraction. Identify the two numbers as *individual observations* from two randomly varying quantities, however, and the character of the problem changes significantly: determining—with any *certainty*—which of the random quantities is “larger” and precisely by how much now requires more than mere subtraction. This is the case with our first example.

### 1.1.1 The Problem

A chemical process using catalyst A (process “A”) is being considered as an alternative to the incumbent process using a different catalyst B (process “B”). The decision in favor of one process over the other is to be based on a comparison of the yield  $Y_A$  obtained from the challenger, and  $Y_B$  from the incumbent, in conjunction with the following economic considerations :

- Achieving target profit objectives for the finished product line requires a process yield *consistently* at or above 74.5%.
- For process A to be a viable alternative, at the barest minimum, its yield must be higher than that for process B.
- Every 1% yield increase over what is currently achievable with the incumbent process translates to significant after tax operating income; however, catalyst A used by the alternative process costs more than catalyst B used by the incumbent process. Including the additional cost of the process modifications required to implement the new technology, *a shift to catalyst A and the new process will be economically viable only if the resulting yield increase exceeds 2%*.

The result of a series of 50 experiments “carefully performed” on each process to determine  $Y_A$  and  $Y_B$  is shown in Table 1.1. Given only the supplied data, what should the Vice President/General Manager of this business do: authorize a switch to the new process A or stay with the incumbent process?

### Mathematical Formulation

Observe that solving this problem requires finding appropriate answers to the following mathematical questions:

1. Is  $Y_A \geq 74.5$ , and  $Y_B \geq 74.5$ , consistently?
2. Is  $Y_A > Y_B$ ?

**TABLE 1.1:** Yield Data  
for Process A versus Process B

| $Y_A$ % |       | $Y_B$ % |       |
|---------|-------|---------|-------|
| 74.04   | 75.29 | 75.75   | 68.41 |
| 75.63   | 75.92 | 74.19   | 68.10 |
| 77.21   | 75.07 | 68.10   | 69.23 |
| 74.23   | 74.92 | 70.14   | 69.23 |
| 76.58   | 77.77 | 74.17   | 70.24 |
| 75.05   | 74.90 | 70.09   | 71.91 |
| 75.69   | 75.31 | 72.63   | 78.41 |
| 75.19   | 77.93 | 71.16   | 73.37 |
| 75.37   | 74.78 | 70.27   | 73.64 |
| 74.47   | 72.99 | 75.82   | 74.42 |
| 73.99   | 73.32 | 72.14   | 78.49 |
| 74.90   | 74.88 | 74.88   | 76.33 |
| 75.78   | 79.07 | 70.89   | 71.07 |
| 75.09   | 73.87 | 72.39   | 72.04 |
| 73.88   | 74.23 | 74.94   | 70.02 |
| 76.98   | 74.85 | 75.64   | 74.62 |
| 75.80   | 75.22 | 75.70   | 67.33 |
| 77.53   | 73.99 | 72.49   | 71.71 |
| 72.30   | 76.56 | 69.98   | 72.90 |
| 77.25   | 78.31 | 70.15   | 70.14 |
| 75.06   | 76.06 | 74.09   | 68.78 |
| 74.82   | 75.28 | 72.91   | 72.49 |
| 76.67   | 74.39 | 75.40   | 76.47 |
| 76.79   | 77.57 | 69.38   | 75.47 |
| 75.85   | 77.31 | 71.37   | 74.12 |

3. If yes, is  $Y_A - Y_B > 2$ ?

Clearly, making the proper decision hinges on our ability to answer these questions with confidence.

### 1.1.2 The Essence of the Problem

Observe that the real essence of the problem is *random variability*: if each experiment had resulted in the same single, constant number for  $Y_A$  and another for  $Y_B$ , the problem would be “deterministic” in character, and each of the 3 associated questions would be trivial to answer. Instead, the random phenomena inherent in the experimental determination of the “true” process yields have been manifested in the observed variability, so that we are uncertain about the “true” values of  $Y_A$  and  $Y_B$ , making it not quite as trivial to solve the problem.

The sources of variability in this case can be shown to include the measurement procedure, the measurement device itself, raw materials, and process conditions. The observed variability is therefore intrinsic to the problem and cannot be idealized away. There is no other way to solve this problem rationally without dealing directly with the random variability.

Next, note that  $Y_A$  and  $Y_B$  data (observations) take on values on a “continuous scale” i.e. yield values are “real” and can be located anywhere on the real line, as opposed to quantities that can take on integer values only (as is the case with the second example discussed later). The variables  $Y_A$  and  $Y_B$  are therefore said to be “continuous” and this example illustrates decision-making under uncertainty when the random phenomena in question involve “continuous variables.”

The main issues with this problem are as follows:

1. *Characterization*: How should the quantities  $Y_A$  and  $Y_B$  be characterized so that the questions raised above can be answered properly?
2. *Quantification*: Are there such things as “true values” of the quantities  $Y_A$  and  $Y_B$ ? If so, how should these “true values” be best quantified?
3. *Application*: How should the characterization and quantification of  $Y_A$  and  $Y_B$  be used to answer the 3 questions raised above?

### 1.1.3 Preliminary “Intuitive” Notions

Before outlining procedures for solving this problem, it is helpful to entertain some notions that the intuition of a good scientist or engineer will suggest. For instance, the concept of the “arithmetic average” of a collection of  $n$  data points,  $x_1, x_2, x_3, \dots, x_n$ , defined by:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (1.1)$$

is well-known to all scientists and engineers, and the intuitive notion of employing this single computed value to “represent” the data set is almost instinctive. It seems reasonable therefore to consider representing  $Y_A$  with the computed average obtained from the data, i.e.  $\bar{y}_A = 75.52$ , and similarly, representing  $Y_B$  with  $\bar{y}_B = 72.47$ . We may now observe right away that  $\bar{y}_A > \bar{y}_B$ , which now seems to suggest not only that  $Y_A > Y_B$ , but since  $\bar{y}_A - \bar{y}_B = 3.05$ , that the difference in fact exceeds the threshold of 2%.

As intuitively appealing as these arguments might be, they raise some important additional questions:

1. The variability of individual values of the data  $y_{A_i}$  around the average value  $\bar{y}_A = 75.52$  is noticeable; that of  $y_{B_i}$  around the average value  $\bar{y}_B = 72.47$  even more so. How confident then are we about the arguments presented above, and in the implied recommendation to prefer process A to B, based as they are on the computed averages? (For example, there are some 8 values of  $y_{B_i} > \bar{y}_A$ ; what should we make of this fact?)
2. Will it (or should it) matter that

$$\begin{aligned} 72.30 < y_{A_i} &< 79.07 \\ 67.33 < y_{B_i} &< 78.41 \end{aligned} \quad (1.2)$$

so that the observed data are seen to vary over a *range* of yield values that is 11.08 units wide for process B as opposed to 6.77 for A? The averages give no indication of these extents of variability.

3. More fundamentally, is it always a good idea to work with averages? How reasonable is it to characterize the entire data set with the average?
4. If new sets of data are gathered, the new averages computed from them will almost surely differ from the corresponding values computed from the current set of data shown here. Observe therefore that the computed averages  $\bar{y}_A$  and  $\bar{y}_B$  are themselves clearly subject to random variability. How can we then be sure that using averages offers any advantages, since, like the original data, these averages are also not free from random variability?
5. How were the data themselves collected? What does it mean concretely that the 50 experiments were “carefully performed”? Is it possible that the experimental protocols used may have impaired our ability to answer the questions posed above adequately? Conversely, are there protocols that are particularly calibrated to improve our ability to answer these questions adequately?

Obviously therefore there is a lot more to dealing with this example problem than merely using the intuitively appealing notion of averages.

Let us now consider a second, different but somewhat complementary, example.

**TABLE 1.2:** Number of “inclusions” on sixty 1-sq meter glass sheets

|   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 2 | 2 |
| 2 | 0 | 2 | 2 | 3 | 2 | 0 | 0 | 2 | 0 |
| 1 | 2 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 |
| 1 | 1 | 5 | 2 | 0 | 0 | 1 | 4 | 1 | 1 |
| 2 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 |
| 1 | 0 | 0 | 2 | 4 | 0 | 1 | 1 | 0 | 1 |

## 1.2 Quality Assurance in a Glass Sheet Manufacturing Process

A key measure of product quality in a glass sheet manufacturing process is the optical attribute known as “inclusions”—particulate flaws (of size exceeding  $0.5 \mu\text{m}$ ) “included” in an otherwise perfectly clear glass sheet. While it is all but inevitable to find *inclusions* in some products, the best manufacturers produce remarkably few glass sheets with these imperfections; and even then, the actual number of *inclusions* on these imperfect sheets is itself usually very low, perhaps 1 or 2.

The specific example in question involves a manufacturer of 1 sq. meter sheets of glass used for various types of building windows. Prior to shipping a batch of manufactured product to customers, a sample of glass sheets from the batch is sent to the company’s Quality Control (QC) laboratory where an optical scanning device is used to determine  $X$ , the number of *inclusions* in each square-meter sheet. The results for 60 samples from a particular batch is shown in Table 1.2.

This particular set of results caused the supervising QC engineer some concern for the following reasons:

1. Historically, the manufacturing process hardly ever produces sheets with more than 3 *inclusions* per square meter; this batch of 60 has three such sheets: two with 4 *inclusions*, and one with 5.
2. Each 1 sq-m sheet with 3 or fewer *inclusions* is acceptable and can be sold to customers unconditionally; sheets with 4 *inclusions* are marginally acceptable so long as a batch of 1000 sheets does not contain more than 20 such sheets; a sheet with 5 or more inclusions is unacceptable and cannot be shipped to customers. All such sheets found by a customer are sent back to the manufacturer (at the manufacturer’s expense) for a full refund. The specific sheet of this type contained in this sample of 60 must therefore be found and eliminated.
3. More importantly, the manufacturing process was designed such that

when operated properly, there will be no more than 3 unacceptable sheets with 5 or more *inclusions* in each batch of 1000 sheets. The process will be uneconomical otherwise.

The question of interest is this: *Does the QC engineer have a reason to be concerned?* Or, stated mathematically, if  $X^*$  is the design value of the number of inclusions per sq. m. associated with the sheets produced by this process, is there evidence in this sample data that  $X > X^*$  so that steps will have to be taken to identify the source of this process performance degradation and then to rectify the problem in order to improve the process performance?

As with the first problem, the primary issue here is also the randomness associated with the variable of interest,  $X$ , the number of *inclusions* per square meter of each glass sheet. The value observed for  $X$  in the QC lab is a randomly varying quantity, not fixed and deterministic. In this case, however, there is little or no contribution to the observed variability from the measurement device: these particulate *inclusions* are relatively few in number and are easily counted *without error* by the optical device. The variability in raw material characteristics, and in particular the control system's effectiveness in maintaining the process conditions at desired values (in the face of inevitable and unpredictable disturbances to the process) all contribute to whether or not there are imperfections, how many there are per square meter, and where they are located on the sheet. Some sheets come out flawless while others end up with a varying number of *inclusions* that cannot be predicted precisely *a priori*. Thus, once again, the observed variability must be dealt with directly because it is intrinsic to the problem and cannot be idealized away.

Next, note that the data in Table 1.2, being counts of distinct entities, take on integer values. The variable  $X$  is therefore said to be "discrete," so that this example illustrates decision-making when the random phenomena in question involve "discrete variables."

---

### 1.3 Outline of a Systematic Approach

Even though the two illustrative problems presented above are different in so many ways (one involves continuous variables, the other a discrete variable; one is concerned with comparing two entities to each other, the other pits a single set of data against a design target), the systematic approach to solving such problems provided by probability and statistics applies to both in a unified way. The fundamental issues at stake may be stated as follows:

*In light of its defining characteristics of intrinsic variability, how should randomly varying quantities be characterized and quantified precisely in order to facilitate the solution of practical problems involving them?*

**TABLE 1.3:** Group classification and frequencies for  $Y_A$  data (from the proposed process)

| $Y_A$ group | Frequency | Relative Frequency |
|-------------|-----------|--------------------|
| 71.51-72.50 | 1         | 0.02               |
| 72.51-73.50 | 2         | 0.04               |
| 73.51-74.50 | 9         | 0.18               |
| 74.51-75.50 | 17        | 0.34               |
| 75.51-76.50 | 7         | 0.14               |
| 76.51-77.50 | 8         | 0.16               |
| 77.51-78.50 | 5         | 0.10               |
| 78.51-79.50 | 1         | 0.02               |
| TOTAL       | 50        | 1.00               |

What now follows is a somewhat informal examination of the ideas and concepts behind these time-tested techniques. The purpose is to motivate and provide context for the more formal discussions in upcoming chapters.

### 1.3.1 Group Classification and Frequency Distributions

Let us revisit the example data sets and consider the following alternative approach to the data representation. Instead of focusing on individual observations as presented in the tables of raw data, what if we sub-divided the observations into small groups (called “bins”) and re-organized the raw data in terms of how frequently members of each group occur? One possible result is shown in Tables 12.3 and 1.4 respectively for process A and process B. (A different bin size will lead to a slightly different group classification but the principles remain the same.)

This reclassification indicates, for instance, that for  $Y_A$ , there is only one observation between 71.51 and 72.50 (the actual number is 72.30), but there are 17 observations between 74.51 and 75.50; for  $Y_B$  on the other hand, 3 observations fall in the [67.51-68.50] group whereas there are 8 observations between 69.51 and 70.50. The “relative frequency” column indicates what *proportion* of the original 50 data points are found in each group. A plot of this reorganization of the data, known as the *histogram*, is shown in Figure 12.8 for  $Y_A$  and Figure 1.2 for  $Y_B$ .

The histogram, a term first used by Pearson in 1895, is a graphical representation of data from a group-classification and frequency-of-occurrence perspective. Each bar represents a distinct group (or class) within the data set, with the bar height proportional to the group frequency. Because this graphical representation provides a picture of how the data are “distributed” in terms of the frequency of occurrence of each group (how much each group

**TABLE 1.4:** Group classification and frequencies for  $Y_B$  data (from the incumbent process)

| $Y_B$ group | Frequency | Relative Frequency |
|-------------|-----------|--------------------|
| 66.51-67.50 | 1         | 0.02               |
| 67.51-68.50 | 3         | 0.06               |
| 68.51-69.50 | 4         | 0.08               |
| 69.51-70.50 | 8         | 0.16               |
| 70.51-71.50 | 4         | 0.04               |
| 71.51-72.50 | 7         | 0.14               |
| 72.51-73.50 | 4         | 0.08               |
| 73.51-74.50 | 6         | 0.12               |
| 74.51-75.50 | 5         | 0.10               |
| 75.51-76.50 | 6         | 0.12               |
| 76.51-77.50 | 0         | 0.00               |
| 77.51-78.50 | 2         | 0.04               |
| 78.51-79.50 | 0         | 0.00               |
| TOTAL       | 50        | 1.00               |

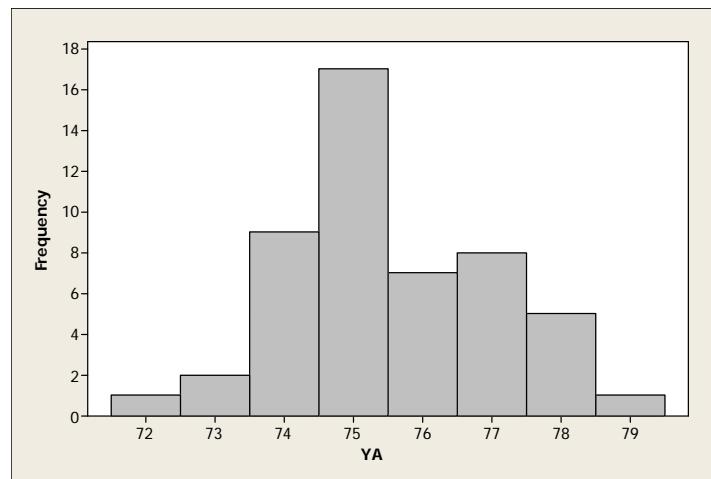


FIGURE 1.1: Histogram for  $Y_A$  data

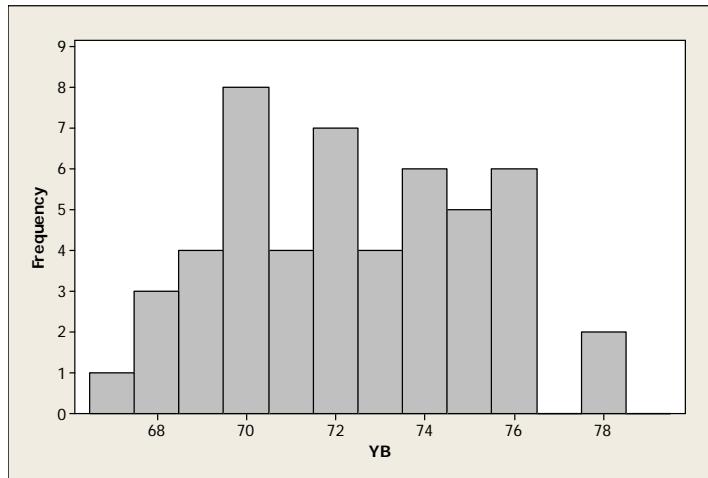


FIGURE 1.2: Histogram for  $Y_B$  data

contributes to the data set), it is often referred to as a *frequency distribution* of the data.

A key advantage of such a representation is how clearly it portrays the nature of the variability associated with each variable. For example, we easily see from Fig 12.8 that the “center of action” for the  $Y_A$  data is somewhere around the group whose bar is centered around 75 (i.e. in the interval [74.51, 75.50]). Furthermore, most of the values of  $Y_A$  cluster in the 4 central groups centered around 74, 75, 76 and 77. In fact, 41 out of the 50 observations, or 82%, fall into these 4 groups; groups further away from the center of action (to the left as well as to the right) contribute less to the  $Y_A$  data. Similarly, Fig 1.2 shows that the “center of action” for the  $Y_B$  data is located somewhere around the group in the [71.51, 72.50] interval but it is not as sharply defined as it was with  $Y_A$ . Also the values of  $Y_B$  are more “spread out” and do not cluster as tightly around this central group.

The histogram also provides *quantitative* insight. For example, we see that 38 of the 50  $Y_A$  observations (or 76%) are greater than 74.51; only 13 out of the 50  $Y_B$  observations (or 26%) fall into this category. Also, exactly 0% of  $Y_A$  observations are less than or equal to 71.50 compared with 20 out of 50 observations (or a staggering 40%) of  $Y_B$  observations. Thus, if these data sets can be considered as representative of the overall performance of each process, then it is reasonable to conclude, for example, that there is a better chance of obtaining yields greater than 74.50 with process A than with process B (a 76% chance compared to a 26% chance). Similarly, while it is highly unlikely that process A will ever return yields less than 71.50, there is a not-insignificant chance (40%) that the yield obtained from process B will be less than 71.50. What is thus beginning to emerge are the faint outlines of

**TABLE 1.5:** Group classification and frequencies for the *inclusions* data

| $X$   | Frequency | Relative Frequency |
|-------|-----------|--------------------|
| 0     | 22        | 0.367              |
| 1     | 23        | 0.383              |
| 2     | 11        | 0.183              |
| 3     | 1         | 0.017              |
| 4     | 2         | 0.033              |
| 5     | 1         | 0.017              |
| 6     | 0         | 0.000              |
| TOTAL | 60        | 1.000              |

a rigorous framework for characterizing and quantifying random variability, with the histogram providing this first glimpse.

It is important to note that the advantage provided by the histogram comes at the expense of losing the *individuality* of each observation. Having gone from 50 raw observations each to 8 groups for  $Y_A$ , and a slightly larger 12 groups for  $Y_B$ , there is clearly a loss of resolution: the individual identities of the original observations are no longer visible *from the histogram*. (For example, the identities of each of the 17  $Y_A$  observations that make up the group in the interval [74.51,75.50] have been melded into that of a single, monolithic bar in the histogram.) But this is not necessarily a bad thing. As we demonstrate in upcoming chapters, a fundamental tenet of the probabilistic approach to dealing with randomly varying phenomena is an abandonment of the individual observation as the basis for theoretical characterization, in favor of an ensemble description. For now, it suffices to be able to see from this example that the clarity with which the histogram portrays data variability has been achieved by trading off the individual observation's identity for the ensemble identity of groups. But keep in mind that what the histogram offers is simply an alternative (albeit more informative) way of representing the same identical information contained in the data tables.

Let us now return to the second problem. In this case, the group classification and frequency distribution for the raw *inclusions* data is shown in Table 1.5. Let it not be lost on the reader that while the groups for the yield data sets were created from intervals of finite length, no such *quantization* is necessary for the *inclusions* data since in this case, the variable of interest,  $X$ , is naturally discrete. This fundamental difference between continuous variables (such as  $Y_A$  and  $Y_B$ ) and discrete variables (such as  $X$ ) will continue to surface at various stages in subsequent discussions.

The histogram for the *inclusions* data is shown in Fig 1.3 where several characteristics are now clear: for example, 75% of the glass sheets (45 out of 60) are either perfect or have only a single (almost inconsequential) inclusion; only

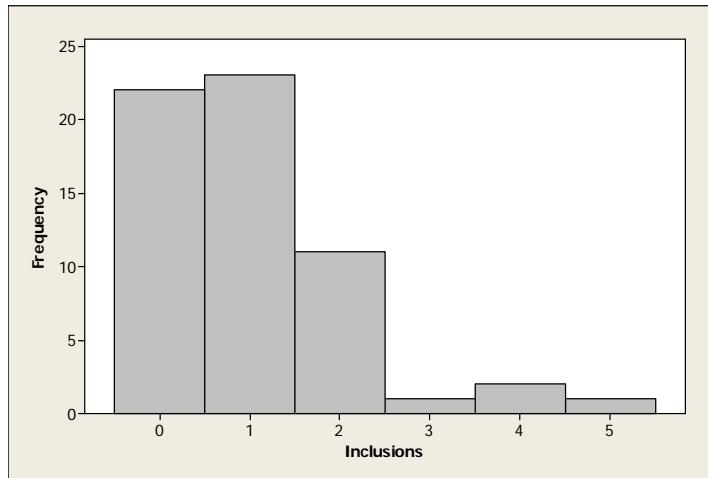


FIGURE 1.3: Histogram of inclusions data

5% of the glass sheets (3 out of 60) have more than 3 inclusions, the remaining 95% have 3 or fewer; 93.3% (56 out of 60) have 2 or fewer inclusions. The important point is that such quantitative characteristics of the data variability (made possible by the histogram) is potentially useful for answering practical questions about what one can reasonably expect from this process.

### 1.3.2 Theoretical Distributions

How can the benefits of the histogram be consolidated into a useful tool for quantitative analysis of randomly varying phenomena? The answer: by appealing to a fundamental axiom of random phenomena: that conceptually, as more observations are made, the shape of the data histogram stabilizes, and tends to the form of the *theoretical distribution* that characterizes the random phenomenon in question, *in the limit as the total number of observations approaches infinity*. It is important to note that this concept does not necessarily require that an infinite number of observations actually be obtained in practice, even if this were possible. The essence of the concept is that an underlying theoretical distribution exists for which the frequency distribution represented by the histogram is but a “finite sample” approximation; that the underlying theoretical distribution is an ideal model of the particular phenomenon responsible for generating the finite number of observations contained in the current data set; and hence that this theoretical distribution provides a reasonable mathematical characterization of the random phenomenon.

As we show later, these theoretical distributions may be derived from first principles given sufficient knowledge regarding the underlying random phenomena. And, as the brief informal examination of the illustrative histograms

above indicates, these theoretical distributions can be used for various things. For example, even though we have not yet provided any concrete definition of the term *probability*, neither have we given any concrete justifications of its usage in this context, still from the discussion in the previous section, the reader can intuitively attest to the reasonableness of the following statements: “the probability that  $Y_A \geq 74.5$  is  $\approx 0.76$ ”; or “the probability that  $Y_B \geq 74.5$  is  $\approx 0.26$ ”; or “the probability that  $X \leq 1$  is  $\approx 0.75$ ”. Parts II and III are devoted to establishing these ideas more concretely and more precisely.

## A Preview

It turns out that the theoretical distribution for each yield data set is:

$$f(y|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(y-\mu)^2}{2\sigma^2}}; -\infty < y < \infty \quad (1.3)$$

which, when superimposed on each histogram, is shown in Fig 1.4 for  $Y_A$ , and Fig 1.5 for  $Y_B$ , when the indicated “characteristic parameters” are specified as  $\mu = 75.52, \sigma = 1.43$  for  $Y_A$ , and  $\mu = 72.47, \sigma = 2.76$  for  $Y_B$ .

Similarly, the theoretical distribution for the *inclusions* data is:

$$f(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}; x = 0, 1, 2, \dots \quad (1.4)$$

where the characteristic parameter  $\lambda = 1.02$  is the average number of *inclusions* in each glass sheet. In similar fashion to Eq 4.155, it also provides a theoretical characterization and quantification of the random phenomenon responsible for the variability observed in the *inclusions* data. From it we are able, for example, to compute the “theoretical probabilities” of observing  $0, 1, 2, \dots$ , *inclusions* in any one glass sheet manufactured by this process. A plot of this theoretical probability distribution function is shown in Fig 22.41 (compare with the histogram in Fig 1.3).

The full detail of precisely what all this means is discussed in subsequent chapters; for now, this current brief preview serves the purpose of simply indicating how the expression in Eqs 4.155 and 4.40 provide a theoretical means of characterizing (and quantifying) the random phenomenon involved respectively in the yield data and in the inclusions data. Expressions such as this are called “probability distribution functions” (pdfs) and they provide the basis for rational analysis of random variability via the concept of “probability.” Precisely what this concept of “probability” is, how it gives rise to pdfs, and how pdfs are used to solve practical problems and provide answers to the sorts of questions posed by these illustrative examples, constitute the primary focus of the remaining chapters in the book.

At this point, it is best to defer the rest of the discussion until when we revisit these two problems at appropriate places in upcoming chapters where we show that:

1.  $Y_A$  indeed may be considered as “greater” than  $Y_B$ , and in particular, that  $Y_A - Y_B > 2$ , up to a specific, quantifiable “degree of confidence,”

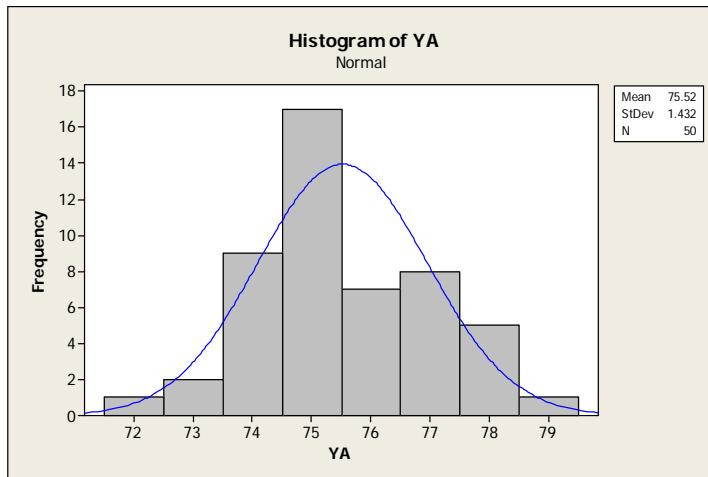


FIGURE 1.4: Histogram for  $Y_A$  data with superimposed theoretical distribution

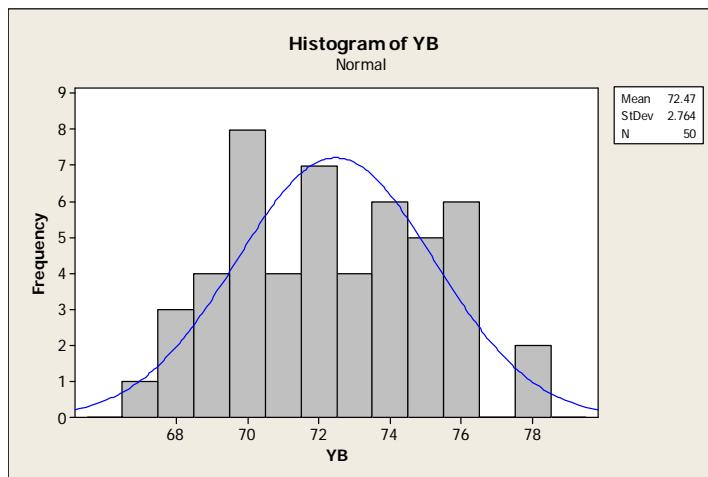
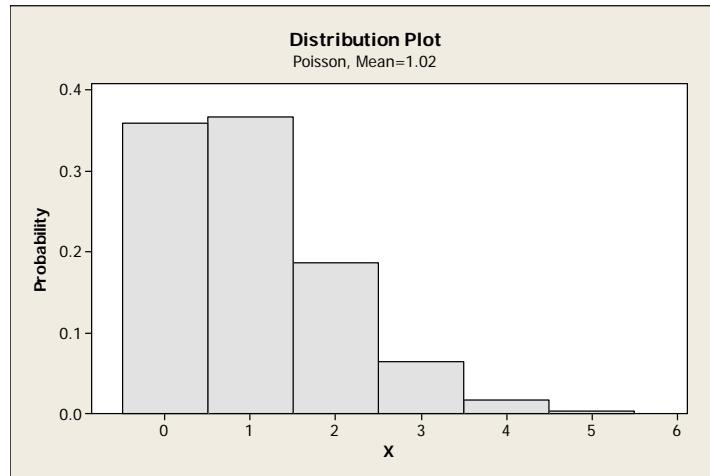


FIGURE 1.5: Histogram for  $Y_B$  data with superimposed theoretical distribution



**FIGURE 1.6:** Theoretical probability distribution function for a Poisson random variable with parameter  $\lambda = 1.02$ . Compare with the inclusions data histogram in Fig 1.3

2. There is in fact no evidence in the *inclusions* data to suggest that the process has deviated from its design target; i.e. that there is no reason to believe that  $X \neq X^*$ , again up to a specific, quantifiable “degree of confidence.”

## 1.4 Summary and Conclusions

We have introduced two practical problems in this chapter to illustrate the complications caused by the presence of randomly varying phenomena in engineering problems. One problem involved determining which of two *continuous* variables is larger; the other involved determining if a *discrete* variable has deviated from its design target. Without the presence of random variability, each problem would ordinarily have been trivial to solve. However, with intrinsic variability that could not be idealized away, it became clear that special techniques capable of coping explicitly with randomly varying phenomena would be required to solve these problems satisfactorily. We did not solve the problems, of course (that is reserved for later); we simply provided an outline of a systematic approach to solving them, which required introducing some concepts that are to be explored fully later. As a result, the very brief introduction of the frequency distribution, the graphical histogram, and the theoretical distribution function was intended to serve merely as a preview of

upcoming detailed discussions concerning how randomly varying phenomena are analyzed systematically.

Here are some of the main points of the chapter again:

- The presence of random variability often complicates otherwise straightforward problems so that specialized solution techniques are required;
- Frequency distributions and histograms provide a particularly informative perspective of random variations intrinsic to experimental data;
- The probability distribution function — the theoretical limit to which the frequency distribution (and histogram) tends — provides the basis for systematic analysis of randomly varying phenomena.

## REVIEW QUESTIONS

1. What decision is to be made in the yield improvement problem of Section 1.1?
2. What are the economic factors to be taken into consideration in deciding what to do with the yield improvement problem?
3. What is the essence of the yield improvement problem as discussed in Section 1.1?
4. What are some of the sources of variability associated with the process yields?
5. Why are the yield variables,  $Y_A$  and  $Y_B$ , *continuous* variables?
6. What single value is suggested as “intuitive” for representing a collection of  $n$  data points,  $x_1, x_2, \dots, x_n$ ?
7. What are some of the issues raised by entertaining the idea of representing the yield data sets with the arithmetic averages  $\bar{y}_A$  and  $\bar{y}_B$ ?
8. Why is the number of *inclusions* found on each glass sheet a discrete variable?
9. What are some sources of variability associated with the glass manufacturing process which may ultimately be responsible for the variability observed in the number of *inclusions*?
10. What is a frequency distribution and how is it obtained from raw data?
11. Why will bin size affect the appearance of a frequency distribution?
12. What is a histogram and how is it obtained from data?
13. What is the primary advantage of a histogram over a table of raw data?

- 14.** What is the relationship between a histogram and a theoretical distribution?
- 15.** What are the expressions in Eqs (4.155) and (4.40) called? These equations provide the basis for what?

## EXERCISES

### Section 1.1

- 1.1** The variance of a collection of  $n$  data points,  $y_1, y_2, \dots, y_n$ , is defined as:

$$s^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1} \quad (1.5)$$

where  $\bar{y}$  is the arithmetic average of the data set. From the yield data in Table 1.1, obtain the variances  $s_A^2$  and  $s_B^2$  for the  $Y_A$  and  $Y_B$  data sets, respectively. Which is greater,  $s_A^2$  or  $s_B^2$ ?

- 1.2** Even though the data sets in Table 1.1 were not generated in pairs, obtain the 50 differences,

$$d_i = y_{A_i} - y_{B_i}; i = 1, 2, \dots, 50, \quad (1.6)$$

for corresponding values of  $Y_A$  and  $Y_B$  as presented in this table. Obtain a histogram of  $d_i$  and compute the arithmetic average,

$$\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i. \quad (1.7)$$

What do these results suggest about the possibility that  $Y_A$  may be greater than  $Y_B$ ?

- 1.3** A set of theoretical results to be established later (see Chapter 4 Exercises) state that, for  $d_i$  and  $\bar{d}$  defined in Eq (1.7), and variance  $s^2$  defined in Exercise 1,

$$\bar{d} = \bar{y}_A - \bar{y}_B \quad (1.8)$$

$$s_d^2 = s_A^2 + s_B^2 \quad (1.9)$$

Confirm these results specifically for the data in Table 1.1.

### Section 1.2

- 1.4** From the data in Table 1.2, obtain  $s_x^2$ , the variance of the inclusions.

- 1.5** The random variable,  $X$ , representing the number of inclusions, is purported to be a Poisson random variable (see Chapter 8). If true, then the average,  $\bar{x}$ , and variance,  $s_x^2$ , are theoretically equal. Compare the values computed for these two quantities from the data set in Table 1.2. What do these results suggest about the possibility that  $X$  may in fact be a Poisson random variable?

### Section 1.3

- 1.6** Using a bin size of 0.75, obtain relative frequencies for  $Y_A$  and  $Y_B$  data and the corresponding histograms. Repeat this exercise for a bin size of 2.0. Compare these

two sets of histograms with the corresponding histograms in Figs 12.8 and 1.2.

**1.7** From the frequency distribution in Table 12.3 and the values computed for the average,  $\bar{y}_A$ , and variance,  $s_A^2$  of the yield data set,  $Y_A$ , determine the percentage of the data contained in the interval  $\bar{y}_A \pm 1.96s_A$ , where  $s_A$  is the positive square root of the variance,  $s_A^2$ .

**1.8** Repeat Exercise 1.7 for the  $Y_B$  data in Table 1.4. Determine the percentage of the data contained in the interval  $\bar{y}_B \pm 1.96s_B$ .

**1.9** From Table 1.5 determine the value of  $x$  such that only 5% of the data exceeds this value.

**1.10** Using  $\mu = 75.52$  and  $\sigma = 1.43$ , compute theoretical values of the function in Eq 4.155 at the center points of the frequency groups for the  $Y_A$  data in Table 12.3; i.e., for  $y = 72, 73, \dots, 79$ . Compare these theoretical values with the corresponding relative frequency values.

**1.11** Repeat Exercise 1.10 for  $Y_B$  data and Table 1.4.

**1.12** Using  $\lambda = 1.02$ , compute theoretical values of the function  $f(x|\lambda)$  in Eq 4.40 at  $x = 0, 1, 2, \dots, 6$  and compare with the corresponding relative frequency values in Table 1.5.

## APPLICATION PROBLEMS

**1.13** The data set in the table below is the time (in months) from receipt to publication (sometimes known as *time-to-publication*) of 85 papers published in the January 2004 issue of a leading chemical engineering research journal.

|      |      |      |      |      |
|------|------|------|------|------|
| 19.2 | 15.1 | 9.6  | 4.2  | 5.4  |
| 9.0  | 5.3  | 12.9 | 4.2  | 15.2 |
| 17.2 | 12.0 | 17.3 | 7.8  | 8.0  |
| 8.2  | 3.0  | 6.0  | 9.5  | 11.7 |
| 4.5  | 18.5 | 24.3 | 3.9  | 17.2 |
| 13.5 | 5.8  | 21.3 | 8.7  | 4.0  |
| 20.7 | 6.8  | 19.3 | 5.9  | 3.8  |
| 7.9  | 14.5 | 2.5  | 5.3  | 7.4  |
| 19.5 | 3.3  | 9.1  | 1.8  | 5.3  |
| 8.8  | 11.1 | 8.1  | 10.1 | 10.6 |
| 18.7 | 16.4 | 9.8  | 10.0 | 15.2 |
| 7.4  | 7.3  | 15.4 | 18.7 | 11.5 |
| 9.7  | 7.4  | 15.7 | 5.6  | 5.9  |
| 13.7 | 7.3  | 8.2  | 3.3  | 20.1 |
| 8.1  | 5.2  | 8.8  | 7.3  | 12.2 |
| 8.4  | 10.2 | 7.2  | 11.3 | 12.0 |
| 10.8 | 3.1  | 12.8 | 2.9  | 8.8  |

(i) Generate a histogram of this data set. Comment on the “shape” of this histogram

and why, from the nature of the variable in question, such a shape may not be surprising.

(ii) From the histogram of the data, what is the “most popular” time-to-publication, and what fraction of the papers took longer than this to publish?

**1.14** Refer to Problem 1.13. Let each raw data entry in the data table be  $x_i$ .

(i) Generate a set of 85 “sample average publication time,”  $y_i$ , from 20 consecutive times as follows:

$$y_1 = \frac{1}{20} \sum_{i=1}^{20} x_i \quad (1.10)$$

$$y_2 = \frac{1}{20} \sum_{i=2}^{21} x_i \quad (1.11)$$

$$y_3 = \frac{1}{20} \sum_{i=3}^{22} x_i \quad (1.12)$$

$$\dots = \dots$$

$$y_j = \frac{1}{20} \sum_{i=j}^{20+(j-1)} x_i \quad (1.13)$$

For values of  $j \geq 66$ ,  $y_j$  should be obtained by replacing  $x_{86}, x_{87}, x_{88}, \dots$ , which do not exist, with  $x_1, x_2, x_3, \dots$ , respectively (i.e., for these purposes treat the given  $x_i$  data like a “circular array”). Plot the histogram for this generated  $y_i$  data and compare the shape of this histogram with that of the original  $x_i$  data.

(ii) Repeat part (i) above, this time for  $z_i$  data generated from:

$$z_j = \frac{1}{20} \sum_{i=j}^{20+(j-1)} y_i \quad (1.14)$$

for  $j = 1, 2, \dots, 85$ . Compare the histogram of the  $z_i$  data with that of the  $y_i$  data and comment on the effect of “averaging” on the shape of the data histograms.

**1.15** The data shown in the table below is a four-year record of the number of “recordable” safety incidents occurring at a plant site each month.

|   |   |   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 1 | 0 | 1 |
| 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2 | 2 | 0 | 1 | 2 | 0 | 1 | 2 | 1 | 1 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |

(i) Find the average number of safety incidents per month and the associated variance. Construct a frequency table of the data and plot a histogram.

(ii) From the frequency table and the histogram, what can you say about the “chances” of obtaining each of the following observations, where  $x$  represents the number of observed safety incidents per month:  $x = 0, x = 1, x = 2, x = 3, x = 4$  and  $x = 5$ ?

(iii) Consider the postulate that a reasonable model for this phenomenon is:

$$f(x) = \frac{e^{-0.5} 0.5^x}{x!} \quad (1.15)$$

where  $f(x)$  represents the theoretical “probability” of recording exactly  $x$  safety incidents per month. How well does this model fit the data?

- (iv) Assuming that this is a reasonable model, discuss how you would use it to answer the question: *If, over the most recent four-month period, the plant recorded 1, 3, 2, 3 safety incidents respectively, is there evidence that there has been a “real increase” in the number of safety incidents?*

**1.16** The table below shows a record of the “before” and “after” weights (in pounds) of 20 patients enrolled in a clinically-supervised ten-week weight-loss program.

| Patient #       | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  |
|-----------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Before Wt (lbs) | 272 | 319 | 253 | 325 | 236 | 233 | 300 | 260 | 268 | 276 |
| After Wt (lbs)  | 263 | 313 | 251 | 312 | 227 | 227 | 290 | 251 | 262 | 263 |
| Patient #       | 11  | 12  | 13  | 14  | 15  | 16  | 17  | 18  | 19  | 20  |
| Before Wt (lbs) | 215 | 245 | 248 | 364 | 301 | 203 | 197 | 217 | 210 | 223 |
| After Wt (lbs)  | 206 | 235 | 237 | 350 | 288 | 195 | 193 | 216 | 202 | 214 |

Let  $X_B$  represent the “Before” weight and  $X_A$  the “After” weight.

- (i) Using the same bin size for each data set, obtain histograms for the  $X_B$  and  $X_A$  data and plot both on the same graph. Strictly on the basis of a visual inspection of these histograms, what can you say about the effectiveness of the weight-loss program in achieving its objective of assisting patients to lose weight?  
(ii) Define the difference variable,  $D = X_B - X_A$ , and from the given data, obtain and plot a histogram for this variable. Again, strictly from a visual inspection of this histogram, what can you say about the effectiveness of the weight-loss program?

**1.17** The data shown in the following table is from an “Assisted Reproductive Technologies” clinic where a cohort of 100 patients under the age of 35 years (the “Younger” group), and another cohort, 35 years and older (the “Older” group), each received five embryos in an in-vitro fertilization (IVF) treatment cycle.

| $x$<br>No. of live<br>births in a<br>delivered<br>pregnancy | $y_O$<br>Total no. of<br>“older patients”<br>(out of 100)<br>with pregnancy outcome $x$ | $y_Y$<br>Total no. of<br>“younger patients”<br>(out of 100)<br>with pregnancy outcome $x$ |
|---|---|---|
| 0   | 32  | 8   |
| 1   | 41  | 25  |
| 2   | 21  | 35  |
| 3   | 5   | 23  |
| 4   | 1   | 8   |
| 5   | 0   | 1   |

The data shows  $x$ , the number of live births per delivered pregnancy, along with how many in each group had the pregnancy outcome of  $x$ . For example, the first entry indicates that the IVF treatment was unsuccessful for 32 of the “older” patients, with the corresponding number being 8 for the “younger” patients; 41 “older” patients delivered singletons, compared with 25 for the younger patients; 21 older patients and 35 younger patients each delivered twins; etc. Obtain a *relative* frequency distribution for these data sets and plot the corresponding histograms. Determine the average number of live births per delivered pregnancy for each group

and compare these values. Comment on whether or not these data sets indicate that the outcomes of the IVF treatments are different for these two groups.



# Chapter 2

## *Random Phenomena, Variability and Uncertainty*

|       |   |    |
|-------|---|----|
| 2.1   | Two Extreme Idealizations of Natural Phenomena .....          | 34 |
| 2.1.1 | Introduction .....  | 34 |
| 2.1.2 | A Chemical Engineering Illustration .....                     | 35 |
|       | Determinism and the PFR .....                                 | 35 |
|       | Randomness and the CSTR .....                                 | 37 |
|       | Theoretical Analysis of the Ideal CSTR's Residence Time ..... | 37 |
| 2.2   | Random Mass Phenomena .....                                   | 41 |
| 2.2.1 | Defining Characteristics .....                                | 41 |
| 2.2.2 | Variability and Uncertainty .....                             | 42 |
| 2.2.3 | Practical Problems of Interest .....                          | 42 |
| 2.3   | Introducing Probability .....                                 | 43 |
| 2.3.1 | Basic Concepts .....  | 43 |
| 2.3.2 | Interpreting Probability .....                                | 44 |
|       | Classical (À-Priori) Probability .....                        | 45 |
|       | Relative Frequency (À-Posteriori) Probability .....           | 46 |
|       | Subjective Probability .....                                  | 46 |
| 2.4   | The Probabilistic Framework .....                             | 47 |
| 2.5   | Summary and Conclusions .....                                 | 48 |
|       | REVIEW QUESTIONS .....  | 49 |
|       | EXERCISES .....   | 50 |
|       | APPLICATION PROBLEMS .....                                    | 51 |

*Through the great beneficence of Providence,  
what is given to be foreseen in the general sphere of masses  
escapes us in the confined sphere of individuals.*

Joannè-Erhard Valentin-Smith (1796–1891)

When John Stuart Mills stated in his 1862 book, *A System of Logic: Ratiocinative and Inductive*, that “...the very events which in their own nature appear most capricious and uncertain, and which in any individual case no attainable degree of knowledge would enable us to foresee, occur, when considerable numbers are taken into account, with a degree of regularity approaching to mathematical ...,” he was merely articulating—astutely for the time—the then-radical, but now well-accepted, concept that randomness in scientific observation is not a synonym for disorder; it is *order of a different kind*. The more familiar kind of order informs “determinism:” the concept that, with sufficient mechanistic knowledge, all physical phenomena are entirely predictable and thus describable by precise mathematical equations. But even classical physics, that archetypal deterministic science, had to make room for this *other* kind

of order when quantum physicists of the 1920's discovered that fundamental particles of nature exhibit irreducible uncertainty (or "chance") in their locations, movements and interactions. And today, most contemporary scientists and engineers are, by training, conditioned to accept both determinism and randomness as intrinsic aspects of the experiential world. The problem, however, is that to many, the basic characteristics of random phenomena and their "order of a different kind" still remain somewhat unfamiliar at a fundamental level.

This chapter is devoted to an expository examination of randomly varying phenomena. Its primary purpose is to introduce the reader to the central characteristic of order-in-the-midst-of-variability, and the sort of analysis this trait permits, *before* diving headlong into a formal study of probability and statistics. The premise of this chapter is that a true appreciation of the nature of randomly varying phenomena at a fundamental level is indispensable to the sort of clear understanding of probability and statistics that will protect the diligent reader from all-too-common misapplication pitfalls.

## 2.1 Two Extreme Idealizations of Natural Phenomena

### 2.1.1 Introduction

In classical physics, the distance,  $x$ , (in meters) traveled in  $t$  seconds by an object launched with an initial velocity,  $u \text{ m/s}$ , and which accelerates at  $a \text{ m/s}^2$ , is known to be given by the expression:

$$x = ut + \frac{1}{2}at^2 \quad (2.1)$$

This is a "deterministic" expression: it consistently and repeatably produces the same result every time identical values of the variables  $u$ ,  $a$ , and  $t$  are specified. The same is true for the expression used in engineering to determine  $Q$ , the rate of heat loss from a house, say in the middle of winter, when the total exposed surface area is  $A \text{ m}^2$ , the inside temperature is  $T_i \text{ K}$ , the outside temperature is  $T_o \text{ K}$ , and the combined heat transfer characteristics of the house walls, insulation, etc., is represented by the so-called "overall heat transfer coefficient"  $U$ ,  $\text{W/m}^2\text{K}$ , i.e.

$$Q = UA(T_i - T_o) \quad (2.2)$$

The rate of heat loss is determined precisely and consistently for any given specific values of each entity on the right hand side of this equation.

The concept of determinism, that the phenomenon in question is *precisely* determinable in every relevant detail, is central to much of science and engineering and has proven quite useful in analyzing real systems, and in solving practical problems—whether it is computing the trajectory of rockets for

launching satellites into orbit, installing appropriate insulation for homes, or designing chemical reactors . However, any assumption of strict determinism in nature is implicitly understood as a convenient idealization resulting from neglecting certain details considered non-essential to the core problem. For example, the capriciousness of the wind and its various and sundry effects have been ignored in Eqs 16.2 and 2.2: no significant wind resistance (or assistance) in the former, negligible convective heat transfer in the latter.

At the other extreme is “randomness,” where the relevant details of the phenomenon in question are indeterminable precisely; repeated observations under identical conditions produce different and randomly varying results; and the observed random variability is essential to the problem and therefore cannot be idealized away. Such is the case with the illustrative problems in Chapter 1 where in one case, the yield obtained from each process may be idealized as follows:

$$y_{Ai} = \eta_A + \epsilon_{Ai} \quad (2.3)$$

$$y_{Bi} = \eta_B + \epsilon_{Bi} \quad (2.4)$$

with  $\eta_A$  and  $\eta_B$  representing the “true” but unknown yields obtainable from processes A and B respectively, and  $\epsilon_{Ai}$  and  $\epsilon_{Bi}$  representing the superimposed randomly varying component—the sources of the random variability evident in each observation  $y_{Ai}$  and  $y_{Bi}$ . Identical values of  $\eta_A$  do not produce identical values of  $y_{Ai}$  in Eq (2.3); neither will identical values of  $\eta_B$  produce identical values of  $y_{Bi}$  in Eq (2.4). In the second case of the glass process and the number of *inclusions* per square meter, the idealization is:

$$x_i = \xi + \epsilon_i \quad (2.5)$$

where  $\xi$  is the “true” number of inclusions associated with the process and  $\epsilon_i$  is the superimposed random component responsible for the observed randomness in the actual number of inclusion  $x_i$  found on each individual glass sheet upon inspection.

These two perspectives, “determinism” and “randomness,” are thus two opposite idealizations of natural phenomena, the former when deterministic aspects of the phenomenon are considered to be overwhelmingly dominant over any random components, the latter case when the random components are dominant and central to the problem. The principles behind each conceptual idealization, and the analysis technique appropriate to each, are now elucidated with a chemical engineering illustration.

### 2.1.2 A Chemical Engineering Illustration

Residence time , the amount of time a fluid element spends in a chemical reactor, is an important parameter in the design of chemical reactors. We wish to consider residence times in two classic reactor configurations: the plug flow reactor (PFR) and the continuous stirred tank reactor (CSTR).

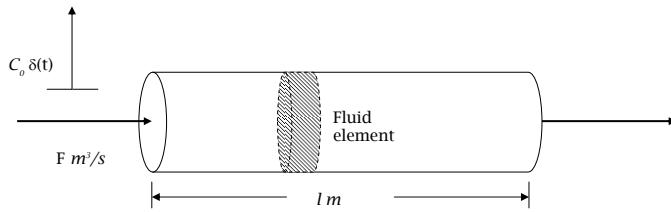


FIGURE 2.1: Schematic diagram of a plug flow reactor (PFR).

### Determinism and the PFR

The plug flow reactor (PFR) is a hollow tube in which reactants that are introduced at one end react as the fluid elements traverse the length of the tube and emerge at the other end. The name comes from the idealization that fluid elements move through as “plugs” with no longitudinal mixing (see Fig 2.1).

The PFR assumptions (idealizations) may be stated as follows:

- the reactor tube ( $l$  m long) has a uniform cross-sectional area,  $A$   $m^2$ ;
- fluid elements move in “plug flow” with a constant velocity,  $v$   $m/s$ , so that the velocity profile is flat;
- the flow rate through the reactor is constant at  $F$   $m^3/s$

Now consider that at time  $t = 0$  we instantaneously inject a bolus of red dye of concentration  $C_0$  moles/ $m^3$  into the inlet stream. The following question is of interest in the study of residence time distributions in chemical reactor design:

How much time does each molecule of red dye spend in the reactor, if we could label them all and observe each one at the reactor exit?

Because of the plug flow idealization, each fluid element moves through the reactor with a constant velocity given by:

$$v = \left( \frac{F}{A} \right) m/s \quad (2.6)$$

and it will take precisely

$$\theta = \frac{l}{v} = \left( \frac{lA}{F} \right) secs \quad (2.7)$$

for each dye element to traverse the reactor. Hence,  $\theta$ , the residence time for an ideal plug flow reactor (PFR) is a deterministic quantity because its value is exactly and precisely determinable from Eq (2.7) given  $F$ ,  $A$  and  $l$ .

Keep in mind that the determinism that informs this analysis of the PFR

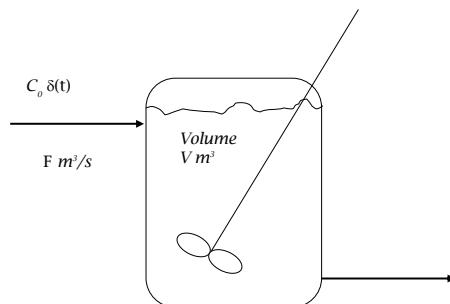


FIGURE 2.2: Schematic diagram of a continuous stirred tank reactor (CSTR).

residence time arises directly as a consequence of the central plug flow idealization. Any departures from such idealization, especially the presence of significant axial dispersion (leading to a non-flat fluid velocity profile), will result in dye molecules no longer arriving at the outlet at precisely the same time.

### Randomness and the CSTR

With the continuous stirred tank reactor (CSTR), the reactant stream continuously flows into a tank that is vigorously stirred to ensure uniform mixing of its content, while the product is continuously withdrawn from the outlet (see Fig 2.2). The assumptions (idealizations) in this case are:

- the reactor tank has a fixed, constant volume,  $V \text{ m}^3$ ;
- the contents of the tank are perfectly mixed.

Once again, let us consider that a bolus of red dye of concentration  $C_0$  moles/ $\text{m}^3$  is instantaneously injected into the inlet stream at time  $t = 0$ ; and again, ask: how much time does each molecule of red dye spend in the reactor? Unlike with the plug flow reactor, observe that it is impossible to answer this question *a-priori*, or precisely: because of the vigorous stirring of the reactor content, some dye molecules will exit almost instantaneously; others will stay longer, some for a *very* long time. In fact, it can be shown that theoretically,  $0 < \theta < \infty$ . Hence in this case,  $\theta$ , the residence time, is a randomly varying quantity that can take on a range of values from 0 to  $\infty$ ; it cannot therefore be adequately characterized as a single number. Notwithstanding, as all chemical engineers know, the random phenomenon of residence times for ideal CSTR's can, and has been, analyzed systematically (see for example, Hill, 1977<sup>1</sup>).

---

<sup>1</sup>C.G. Hill, Jr, *An Introduction to Chemical Engineering Kinetics and Reactor Design*, Wiley, NY, 1977, pp 388-396.

### Theoretical Analysis of the Ideal CSTR's Residence Time

Even though based on chemical engineering principles, the results of the analysis we are about to discuss have fundamental implications for the general nature of the order present in the midst of random variability encountered in other applications, and how such order provides the basis for analysis. (As an added bonus, this analysis also provides a non-probabilistic view of ideas usually considered the exclusive domain of probability).

By carrying out a material balance around the CSTR, (i.e., that the rate of accumulation of mass within a prescribed volume must equal the difference between the rate of input and the rate of output) it is possible to develop a mathematical model for this process as follows: If the volumetric flow rate into and out of the reactor are equal and given by  $F \text{ m}^3/\text{s}$ , if  $C(t)$  represents the molar concentration of the dye in the well-mixed reactor, then by the assumption of perfect mixing, this will also be the dye concentration at the exit of the reactor. The material balance equation is:

$$V \frac{dC}{dt} = FC_{in} - FC \quad (2.8)$$

where  $C_{in}$  is the dye concentration in the inlet stream. If we define the parameter  $\tau$  as

$$\tau = \frac{V}{F} \quad (2.9)$$

and note that the introduction of a bolus of dye of concentration  $C_0$  at  $t = 0$  implies:

$$C_{in} = C_0\delta(t) \quad (2.10)$$

where  $\delta(t)$  is the Dirac delta function, then Eq (2.8) becomes:

$$\tau \frac{dC}{dt} = -C + C_0\delta(t) \quad (2.11)$$

a simple, linear first order ODE whose solution is:

$$C(t) = \frac{C_0}{\tau} e^{-t/\tau} \quad (2.12)$$

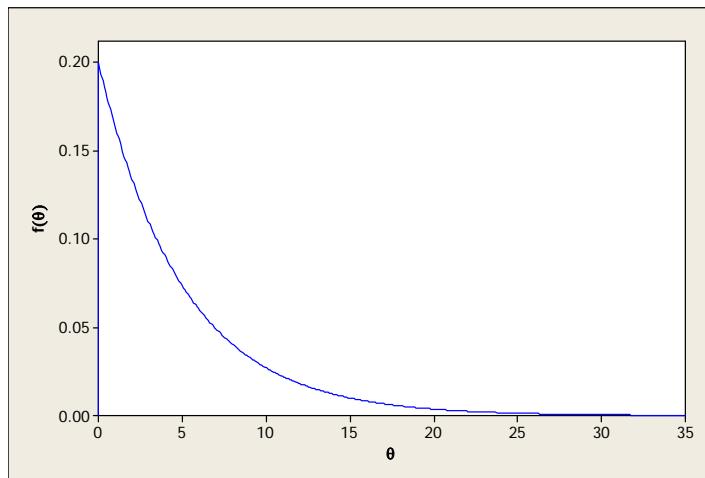
If we now define as  $f(\theta)$ , the *instantaneous* fraction of the initial number of injected dye molecules exiting the reactor at time  $t = \theta$  (those with residence time  $\theta$ ), i.e.

$$f(\theta) = \frac{C(\theta)}{C_0} \quad (2.13)$$

we obtain immediately from Eq (2.12) that

$$f(\theta) = \frac{1}{\tau} e^{-\theta/\tau} \quad (2.14)$$

recognizable to all chemical engineers as the familiar “exponential” *instantaneous* residence time distribution function for the ideal CSTR. The reader



**FIGURE 2.3:** Instantaneous residence time distribution function for the CSTR: (with  $\tau = 5$ ).

should take good note of this expression: it shows up a few more times and in various guises in subsequent chapters. For now, let us observe that, even though (a) the residence time for a CSTR,  $\theta$ , exhibits random variability, potentially able to take on values between 0 and  $\infty$  (and is therefore not describable by a single value); so that (b) it is therefore impossible to determine with absolute certainty precisely when any individual dye molecule will leave the reactor; even so (c) the function,  $f(\theta)$ , shown in Eq (4.41), mathematically characterizes the behavior of the entire ensemble of dye molecules, but in a way that requires some explanation:

1. It represents how the residence times of fluid particles in the well-mixed CSTR are “distributed” over the range of possible values  $0 < \theta < \infty$  (see Fig 2.3).
2. This distribution of residence times is a well-defined, well-characterized function, but it is not a description of the precise amount of time a particular individual dye molecule will spend in the reactor; rather it is a description of how many (or what fraction) of the *entire collection* of dye molecules will spend what amount of time in the reactor. For example, in broad terms, it indicates that a good fraction of the molecules have relatively short residence times, exiting the reactor quickly; a much smaller but non-zero fraction have relatively long residence times. It can also provide more precise statements as follows.
3. From this expression (Eq (4.41)), we can determine the fraction of dye molecules that have remained in the reactor for an amount of time less than or equal to some time  $t$ , (i.e. molecules exiting the reactor with

“age” less than or equal to  $t$ ): we do this by integrating  $f(\theta)$  with respect to  $\theta$ , as follows, to obtain

$$F(t) = \int_0^t \frac{1}{\tau} e^{-\theta/\tau} d\theta = \left(1 - e^{-t/\tau}\right) \quad (2.15)$$

from which we see that  $F(0)$ , the fraction of dye molecules with age less than or equal to zero is exactly zero: indicating the intuitively obvious that, no matter how vigorous the mixing, each dye molecule spends at least a finite, non-zero, amount of time in the reactor (no molecule exits instantaneously upon entry).

On the other hand,  $F(\infty) = 1$ , since

$$F(\infty) = \int_0^\infty \frac{1}{\tau} e^{-\theta/\tau} d\theta = 1 \quad (2.16)$$

again indicating the obvious: if we wait long enough, *all* dye molecules will eventually exit the reactor as  $t \rightarrow \infty$ . In other words, the fraction of molecules exiting the reactor with age less than  $\infty$  is exactly 1.

4. Since the fraction of molecules that will have remained in the reactor for an amount of time less than or equal to  $t$  is  $F(t)$ , and the fraction that will have remained in the reactor for less than or equal to  $t + \Delta t$  is  $F(t + \Delta t)$ , then the fraction with residence time in the infinitesimal interval between  $t$  and  $t + \Delta t$  is given by:

$$\Phi[t \leq \theta \leq (t + \Delta t)] = F(t + \Delta t) - F(t) = \int_t^{t+\Delta t} \frac{1}{\tau} e^{-\theta/\tau} d\theta \quad (2.17)$$

which, for very small  $\Delta t$ , simplifies to:

$$\Phi[t \leq \theta \leq (t + \Delta t)] \approx f(t)\Delta t \quad (2.18)$$

5. And finally, the “average residence time” may be determined from the expression in Eq (4.41) (and Eq (2.16)) as:

$$\bar{\theta} = \frac{\int_0^\infty \frac{1}{\tau} \theta e^{-\theta/\tau} d\theta}{\int_0^\infty \frac{1}{\tau} e^{-\theta/\tau} d\theta} = \frac{\frac{1}{\tau} \int_0^\infty \theta e^{-\theta/\tau} d\theta}{1} = \tau \quad (2.19)$$

where the numerator integral is evaluated via integration by parts. Observe from the definition of  $\tau$  above (in Eq (2.9)) that this result makes perfect sense, strictly from the physics of the problem: particles in a stream flowing at the rate  $F \text{ m}^3/\text{s}$  through a well-mixed reactor of volume  $V \text{ m}^3$ , will spend an average of  $V/F = \tau$  seconds in the reactor.

We now observe in conclusion two important points: (i) even though at no point in the preceding discussion have we made any overt or explicit appeal

to the concepts of probability, the unmistakable fingerprints of probability are evident all over (as upcoming chapters demonstrate concretely, but perhaps already recognizable to those with some familiarity with such concepts); (ii) Nevertheless, this characterizing model in Eq (4.41) was made possible via first-principles knowledge of the underlying phenomenon. This is a central characteristic of random phenomena: that appropriate theoretical characterizations are almost always possible in terms of ideal ensemble models of the observed variability dictated by the underlying phenomenological mechanism.

---

## 2.2 Random Mass Phenomena

### 2.2.1 Defining Characteristics

In such diverse areas as actuarial science, biology, chemical reactors, demography, economics, finance, genetics, human mortality, manufacturing quality assurance, polymer chemistry, etc., one repeatedly encounters a surprisingly common theme whereby phenomena which, on an individual level, appear entirely unpredictable, are well-characterized as ensembles (as demonstrated above with residence time distribution in CSTR's). For example, as far back as 1662, in a study widely considered to be the genesis of population demographics and of modern actuarial science by which insurance premiums are determined today, the British haberdasher, John Graunt (1620-1674), had observed that the number of deaths and the age at death in London were surprisingly predictable for the entire population even though it was impossible to predict which individual would die when and in what manner. Similarly, while the number of monomer molecules linked together in any polymer molecule chain varies considerably, how many chains of a certain length a batch of polymer product contains can be characterized fairly predictably.

Such natural phenomena noted above have come to be known as *Random Mass Phenomena*, with the following defining characteristics:

1. Individual observations appear irregular because it is not possible to predict each one with certainty; but
2. The ensemble or aggregate of all possible outcomes is regular, well-characterized and determinable;
3. The underlying phenomenological mechanisms accounting for the “nature” and occurrence of the specific observations determines the character of the ensemble;
4. Such phenomenological mechanisms may be known mechanistically (as was the case with the CSTR), or its manifestation may only be deter-

mined from data (as was the case with John Graunt's mortality tables of 1662).

This fortunate circumstance—aggregate predictability amidst individual irregularities—is why the primary issue with random phenomena analysis boils down to how to use ensemble descriptions and characterization to carry out systematic analysis of the behavior of individual observations.

### 2.2.2 Variability and Uncertainty

While ensemble characterizations provide a means of dealing systematically with random mass phenomena, many practical problems still involve making decisions about *specific*, inherently unpredictable, outcomes. For example, the insurance company still has to decide what premium to charge each individual on a person-by-person basis. When decisions must be made about specific outcomes of random mass phenomena, uncertainty is an inevitable consequence of the inherent variability. Furthermore, the extent or degree of variability directly affects the degree of uncertainty: tighter clustering of possible outcomes implies less uncertainty, whereas a broader distribution of possible outcomes implies more uncertainty. The most useful mathematical characterization of ensembles must therefore permit not only systematic analysis, but also a rational quantification of the degree of variability inherent in the ensemble, and the resulting uncertainty associated with each individual observation as a result.

### 2.2.3 Practical Problems of Interest

Let  $x_i$  represent individual observations,  $i = 1, 2, \dots, n$ , from a random mass phenomenon; let  $X$  be the actual variable of interest, different and distinct from  $x_i$ , this latter being merely one out of many other possible realizations of  $X$ . For example,  $X$  can be the number of live births delivered by a patient after a round of in-vitro fertilization treatment, a randomly varying quantity; whereas  $x_i = 2$  (i.e. twins) is the specific outcome observed for a specific patient after a specific round of treatment. For now, let the aggregate description we seek be represented as  $f(x)$  (see for example, Eq (4.41) for the CSTR residence time); what this is and how it is obtained is discussed later. In practice, only data in the form of  $\{x_i\}_{i=1}^n$  observations is available. The desired aggregate description,  $f(x)$ , must be understood in its proper context as a descriptor of the (possibly infinite) collection of all possible outcomes of which the observed data is only a “sample.” The fundamental problems of random phenomena analysis may now be stated formally as follows:

1. Given  $\{x_i\}_{i=1}^n$  what can we say about the complete  $f(x)$ ?
2. Given  $f(x)$  what can we say about the specific  $x_i$  values (both the observed  $\{x_i\}_{i=1}^n$  and the yet unobserved)?

Embedded in these questions are the following affiliated questions that arise as a consequence: (a) how was  $\{x_i\}_{i=1}^n$  obtained in (1); will the procedure for obtaining the data affect how well we can answer question 1? (b) how was  $f(x)$  determined in (2)?

Subsequent chapters are devoted to dealing with these fundamental problems systematically and in greater detail.

## 2.3 Introducing Probability

### 2.3.1 Basic Concepts

Consider the prototypical random phenomenon for which the individual observation (or outcome) is not known with certainty *a-priori*, but the complete totality of *all* possible observations (or outcomes) has been (or can be) compiled. Now consider a framework that *assigns* to each individual member of this collection of possible outcomes, a real-valued number between 0 and 1 that represents *the probability of its occurrence*, such that:

1. an outcome that is certain to occur is assigned the number 1;
2. an outcome that is certain *not* to occur is assigned the number 0;
3. any other outcome falling between these two extremes is assigned a number that reflects the extent or degree of certainty (or uncertainty) associated with its occurrence.

Notice how this represents a shift in focus from the individual outcome itself to *the probability of its occurrence*. Using precise definitions and terminology, along with tools of set theory, set functions and real analysis, we show in the chapters in Part II how to develop the machinery for the theory of probability, and the emergence of a compact functional form indicating how the probabilities of occurrence are “distributed” over all possible outcomes. The resulting “probability distribution function” becomes the primary vehicle for analyzing the behavior of random phenomena.

For example, the phenomenon of “inclusions” in manufactured glass sheets discussed in Chapter 1 is well-characterized by the following probability distribution function (pdf) which indicates the probability of observing exactly  $x$  “inclusions” on a glass sheet as

$$f(x) = \frac{e^{-\lambda} \lambda^x}{x!}; x = 0, 1, 2, \dots \quad (2.20)$$

a pdf with a single “parameter,”  $\lambda$ , characteristic of the manufacturing process used to produce the glass sheets. (As shown later,  $\lambda$  is the mean number of

**TABLE 2.1:** Computed probabilities of occurrence of various number of *inclusions* for  $\lambda = 2$  in Eq (9.2)

| $x = \text{No of inclusions}$ | $f(x)$ prob of occurrence |
|-------------------------------|---------------------------|
| 0                             | 0.135                     |
| 1                             | 0.271                     |
| 2                             | 0.271                     |
| 3                             | 0.180                     |
| 4                             | 0.090                     |
| 5                             | 0.036                     |
| :                             | :                         |
| 8                             | 0.001                     |
| $\geq 9$                      | 0.000                     |

“inclusions” on a glass sheet.) Even though we do not know precisely how many “inclusions” will be found on the next glass sheet inspected in the QC lab, given the parameter  $\lambda$ , we can use Eq (9.2) to make statements about the probabilities of individual occurrences. For instance, if  $\lambda = 2$  for a certain process, Eq (9.2) allows us to state that the probability of finding a perfect glass sheet with no “inclusions” in the products made by this process (i.e.  $x = 0$ ) is 0.135; or that the probability of finding 1 “inclusion” is 0.227, coincidentally the same as the probability of finding 2 “inclusions;” or that there is a vanishingly small probability of finding 9 or more “inclusions” in this production facility. The complete set of probabilities computed from Eq (9.2) is shown in Table 2.1.

### 2.3.2 Interpreting Probability

There always seems to be a certain amount of debate over the meaning, definition and interpretation of probability. This is perhaps due to a natural predisposition towards confusing a *conceptual entity* with *how* a numerical value is determined for it. For example, from a certain perspective, “temperature,” as a conceptual entity in Thermodynamics, is a real number *assigned* to an object to indicate its degree of “hotness;” it is distinct from *how* its value is determined (by a thermometer, thermocouple, or any other means). The same is true of “mass,” a quantity assigned in Mechanics to a body to indicate how much matter it contains and how heavy it will be in a gravitational field ; or “distance” assigned in geometry to indicate the closeness of two points in a geometric space. The practical problem of how to determine numerical values for these quantities, even though important in its own right, is a separate issue entirely.

This is how probability should be understood: it is simply a quantity that

is *assigned* to indicate the degree of uncertainty associated with the occurrence of a particular outcome. As with temperature *the conceptual quantity*, how a numerical value is determined for the probability of the occurrence of a particular outcome under any specific circumstance depends on the circumstance itself. To carry the analogy with temperature a bit further: while a thermometer capable of determining temperature to within half a degree will suffice in one case, a more precise device, such as a thermocouple, may be required in another case, and an optical pyrometer for yet another case. Whatever the case, under no circumstance should the *device* employed to determine its numerical value usurp the role of, or become the surrogate for, “temperature” the quantity. This is important in properly interpreting probability, the conceptual entity: how an “appropriate” value is to be determined for probability, an important practical problem in its own right, should not be confused with the quantity itself.

With these ideas in mind, let us now consider several standard perspectives of probability that have evolved over the years. These are best understood as various techniques for *how* numerical values are determined rather than *what* probability is.

### Classical (*À-Priori*) Probability

Consider a random phenomenon for which the total number of possible outcomes is known to be  $N$ , all of which are equally likely; of these, let  $N_A$  be the number of outcomes in which  $A$  is observed (i.e. outcomes that are “favorable” to  $A$ ). Then according to the classical (or *à-priori*) perspective, the probability of the occurrence of outcome  $A$  is defined as

$$P(A) = \frac{N_A}{N} \quad (2.21)$$

For example, in tossing a single perfect die once, the probability of observing a 3 is, according to this viewpoint, evaluated as  $1/6$ , since the total number of possible outcomes is 6 of which only 1 is favorable to the desired observation of 3. Similarly, if  $B$  is the outcome that one observes an odd number of dots, then  $P(B) = 3/6 = 0.5$ .

Observe that according to this view, no experiments have been performed yet; the formulation is based entirely on an *à-priori* enumeration of  $N$  and  $N_A$ . However, this intuitively appealing perspective is not always applicable:

- What if all the outcomes are not equally likely?
- How about random phenomena whose outcomes cannot be characterized as cleanly in this fashion, say, for example, the prospect of a newly purchased refrigerator lasting for 25 years without repair? or the prospect of snow falling on a specific April day in Wisconsin?

What Eq. (2.21) provides is an intuitively appealing (and theoretically sound) means of *determining an appropriate value* for  $P(A)$ ; but it is restricted only

to those circumstances where the random phenomenon in question is characterized in such a way that  $N$  and  $N_A$  are natural and easy to identify.

### Relative Frequency (*À-Posteriori*) Probability

On the opposite end of the spectrum from the *à-priori* perspective is the following alternative: consider an “experiment” that is repeated  $n$  times under identical conditions, where the outcomes involving  $A$  have been observed to occur  $n_A$  times. Then, *à-posteriori*, the probability of the occurrence of outcome  $A$  is defined as

$$P(A) = \lim_{n \rightarrow \infty} \frac{n_A}{n} \quad (2.22)$$

The appeal of this viewpoint is not so much that it is just as intuitive as the previous one, but that it is also empirical, making no assumptions about equal likelihood of outcomes. It is based on the actual performance of “experiments” and the actual *à-posteriori* observation of the relative frequency of occurrences of the desired outcome. This perspective provides a prevalent interpretation of probability as the “theoretical” value of long range relative frequencies. In fact, this is what motivates the notion of the theoretical distribution as the limiting form to which the empirical frequency distribution tends with the acquisition of increasing amounts of data.

However, this perspective also suffers from some limitations:

- How many trials,  $n$ , is sufficient for Eq (2.22) to be useful in practice?
- How about random phenomena for which the desired outcome does not lend itself to repetitive experimentation under identical conditions, say, for example, the prospect of snow falling on a specific April day in Wisconsin? or the prospect of your favorite team winning the basketball championship next year?

Once again, these limitations arise primarily because Eq (2.22) is simply just another means of *determining an appropriate value* for  $P(A)$  that happens to be valid only when the random phenomenon is such that the indicated repeated experimentation is not only possible and convenient, but for which, in practice, truncating after a “sufficiently large” number of trials to produce a finite approximation presents no conceptual dilemma. For example, after tossing a coin 500 times and obtaining 251 heads, declaring that the probability of obtaining a head upon a single toss as 0.5 presents no conceptual dilemma whatsoever.

### Subjective Probability

There is yet another alternative perspective whereby  $P(A)$  is taken simply as a measure of the degree of (personal) belief associated with the postulate that  $A$  will occur, the value having been assigned subjectively by the individual concerned, akin to “betting odds.” Thus, for example, in rolling a perfect die, the probability of obtaining a 3 is assigned strictly on the basis of what the

individual believes to be the likely “odds” of obtaining this outcome, without recourse to enumerating equally likely outcomes (the *à-priori* perspective), or performing the die roll an infinite number of times (the *à-posteriori* perspective).

The obvious difficulty with this perspective is its subjectivity, so that outcomes that are equally likely (on an objective basis) may end up being assigned different probabilities by different individuals. Nevertheless, for those practical applications where the outcomes cannot be enumerated, and for which the experiment cannot be repeated a large number of times, the subjective allocation of probability may be the only viable option, at least *à-priori*. As we show later, it is possible to combine this initial subjective declaration with subsequent limited experimentation in order to introduce objective information contained in data in determining appropriate values of the sought probabilities objectively.

---

## 2.4 The Probabilistic Framework

Beginning with the next chapter, Part II is devoted to an axiomatic treatment of probability, including basic elements of probability theory, random variables, and probability distribution functions, within the context of a comprehensive framework for systematically analyzing random phenomena.

The central conceptual elements of this framework are: (i) a formal representation of uncertain outcomes with the random variable,  $X$ ; and (ii) the mathematical characterization of this random variable by the probability distribution function (pdf),  $f(x)$ . How the probabilities are distributed over the entire aggregate collection of all possible outcomes, expressed in terms of the random variable,  $X$ , is contained in this pdf. The following is a procedure for problem-solving within this framework:

1. *Problem Formulation:* Define and formulate the problem appropriately. Examine the random phenomenon in question, determine the random variable(s), and assemble all available information about the underlying mechanisms;
2. *Model Development:* Identify, postulate, or develop an appropriate ideal model of the relevant random variability in the form of the probability distribution function  $f(x)$ ;
3. *Problem Solution:* Use the model to solve the relevant problem (analysis, prediction, inference, estimation, etc.);
4. *Results validation:* Analyze and validate the result and, if necessary, return to any of the preceding steps as appropriate.

This problem-solving approach is illustrated throughout the rest of the book, particularly in the chapters devoted to actual case studies.

---

## 2.5 Summary and Conclusions

Understanding why, despite appearances, randomly varying phenomena can be subject to analysis of any sort at all is what has occupied our attention in this chapter. Before beginning a formal discussion of random phenomena analysis itself, it was necessary to devote some time to a closer examination of several important foundational issues that are essential to a solid understanding of randomly varying phenomena and their analysis: determinism and randomness; variability and uncertainty; probability and the probabilistic framework for solving problems involving random variability. Using idealized chemical reactors as illustration, we have presented determinism and randomness as two extreme idealizations of natural phenomena. The “residence time” of a dye molecule in the hollow tube of a plug flow reactor (PFR) was used to demonstrate the ideal deterministic variable whose value is fixed and determinable precisely. At the other end of the spectrum is the length of time the dye molecule spends in a vigorously stirred vessel, the ideal continuous stirred tank reactor (CSTR). This time the variable is random and hence impossible to determine precisely *a priori*, but it is not haphazard. The mathematical model derived for the *distribution* of residence times in the CSTR—especially *how* it was obtained from first principles—provides a preview and a chemical engineering analog of what is to come in Chapters 8 and 9, where models are derived for a wide variety of randomly varying phenomena in similar fashion on the basis of underlying phenomenological mechanisms.

We also examined the characteristics of “random mass phenomena,” especially highlighting the co-existence of aggregate predictability in the midst of individual irregularities. This “order-in-the-midst-of-variability” makes possible the use of probability and probability distributions to characterize ensemble behavior mathematically. The subsequent introduction of the *concept* of probability, while qualitative and informal, is nonetheless important. Among other things, it provided a non-technical setting for dealing with the potentially confusing issue of how to interpret probability. In this regard, it bears reiterating that much confusion can be avoided by remembering to keep the *concept* of probability—as a quantity between 0 and 1 used to quantify degree of uncertainty—separate from the means by which numerical values are determined for it. It is in this latter sense that the various interpretations of probability—classical, relative frequency, and subjective—are to be understood: these are all various means of determining a specific value for the probability of a specific outcome; and, depending on the situation at hand, one approach is often more appropriate than others.

Here are some of the main points of the chapter again:

- Randomness does not imply disorder; it is order of a different kind, whereby aggregate predictability co-exists with individual irregularity;
- Determinism and randomness are two extreme idealizations of naturally occurring phenomena, and both are equally subject to rigorous analysis;
- The mathematical framework to be employed in the rest of this book is based on probability, the concept of a random variable,  $X$ , and its mathematical characterization by the pdf,  $f(x)$ .

## REVIEW QUESTIONS

- 1.** If not a synonym for disorder, then what is randomness in scientific observation?
- 2.** What is the concept of determinism?
- 3.** Why are the expressions in Eqs (16.2) and (2.2) considered deterministic?
- 4.** What is an example phenomenon that had to be ignored in order to obtain the deterministic expressions in Eq (16.2)? And what is an example phenomenon that had to be ignored in order to obtain the deterministic expressions in Eq (2.2)?
- 5.** What are the main characteristics of “randomness” as described in Subsection 2.1.1?
- 6.** Compare and contrast determinism and randomness as two opposite idealizations of natural phenomena.
- 7.** Which idealized phenomenon does residence time in a plug flow reactor (PFR) represent?
- 8.** What is the central plug flow idealization in a plug flow reactor, and how will departures from such idealization affect the residence time in the reactor?
- 9.** Which idealized phenomenon does residence time in a continuous stirred-tank reactor (CSTR) represent?
- 10.** On what principle is the mathematical model in Eq (2.8) based?
- 11.** What does the expression in Eq (4.41) represent?
- 12.** What observation by John Graunt is widely considered to be the genesis of population demographics and of modern actuarial science?
- 13.** What are the defining characteristics of random mass phenomena?

- 14.** How does inherent variability give rise to uncertainty?
- 15.** What are the fundamental problems of random phenomena analysis as presented in Subsection 2.2.3?
- 16.** What is the primary mathematical vehicle introduced in Subsection 2.3.1 for analyzing the behavior of random phenomena?
- 17.** What is the classical (*à-priori*) perspective of probability and when is it *not* applicable?
- 18.** What is the relative frequency (*à-posteriori*) perspective of probability and what are its limitations?
- 19.** What is the subjective perspective of probability and under what circumstances is it the only viable option for specifying probability in practice?
- 20.** What are the central conceptual elements of the probabilistic framework?
- 21.** What are the four steps in the procedure for problem-solving within the probabilistic framework?

## EXERCISES

### Section 2.1

**2.1** Solve Eq (2.11) explicitly to confirm the result in Eq (2.12).

**2.2** Plot the expression in Eq (2.15) as a function of the scaled time variable,  $\tilde{t} = t/\tau$ ; determine the percentage of dye molecules with age less than or equal to the mean residence time,  $\tau$ .

**2.3** Show that

$$\int_0^\infty \frac{1}{\tau} \theta e^{-\theta/\tau} d\theta = \tau \quad (2.23)$$

and hence confirm the result in Eq (2.19).

### Section 2.2

**2.4** The following probability distribution functions:

$$f(x) = \frac{1}{3\sqrt{2\pi}} e^{\frac{-x^2}{18}}; -\infty < x < \infty \quad (2.24)$$

and

$$f(y) = \frac{1}{\sqrt{2\pi}} e^{\frac{-y^2}{2}}; -\infty < y < \infty \quad (2.25)$$

represent how the occurrences of all the possible outcomes of the two randomly varying, continuous variables,  $X$  and  $Y$ , are distributed. Plot these two distribution

functions on the same graph. Which of these variables has a higher degree of uncertainty associated with the determination of any particular outcome. Why?

**2.5** When a “fair” coin is tossed 4 times, it is postulated that the probability of obtaining  $x$  heads is given by the probability distribution function:

$$f(x) = \frac{4!}{x!(4-x)!} 0.5^4 \quad (2.26)$$

Determine the probability of obtaining  $x = 0, 1, 2, \dots, 4$  heads. Intuitively, which of these outcomes would you think will be the “most likely?” Are the results of your computation consistent with your intuition?

### Section 2.3

**2.6** In tossing a fair coin once, describe the classical (*a-priori*), relative frequency (*a-posteriori*), and the subjective perspectives of the probability of obtaining a head.

## APPLICATION PROBLEMS

**2.7** For each of the following two-reactor configurations:

- (a) two plug flow reactors in series where the length of reactor 1 is  $l_1$  m, and that of reactor 2 is  $l_2$  m, but both have the same uniform cross-sectional area  $A\ m^2$ ;
- (b) two continuous stirred tank reactors with volumes  $V_1$  and  $V_2\ m^3$ ;
- (c) the PFR in Fig 2.1 followed by the CSTR in Fig 2.2;

given that the flow rate through each reactor ensemble is constant at  $F\ m^3/s$ , obtain the residence time,  $\theta$ , or the residence time distribution,  $f(\theta)$ , as appropriate. Make any assumption you deem appropriate about the concentration  $C_1(t)$  and  $C_2(t)$  in the first and second reactors, respectively.

**2.8** In the summer of 1943 during World War II, a total of 365 warships were attacked by Kamikaze pilots: 180 took evasive action and 60 of these were hit; the remaining 185 counterattacked, of which 62 were hit. Using a relative frequency interpretation and invoking any other assumption you deem necessary, determine the probability that any attacked warship will be hit regardless of tactical response. Also determine the probability that a warship taking evasive action will be hit and the probability that a counterattacking warship will be hit. Compare these three probabilities and discuss what this implies regarding choosing an appropriate tactical response. (A full discussion of this problem is contained in Chapter 7.)

**2.9** Two American National Football League (NFL) teams, **A** and **B**, with respective “Win-Loss” records 9-6 and 12-3 after 15 weeks, are preparing to face each other in the 16<sup>th</sup> and final game of the regular season.

- (i) From a relative frequency perspective of probability, use the supplied information (and any other assumption you deem necessary) to compute the probability of Team A winning any generic game, and also of Team B winning any generic game.

(ii) When the two teams play each other, upon the presupposition that past record is the best indicator of a team's chances of winning a new game, determine reasonable values for  $P(A)$ , the probability that team A wins the game, and  $P(B)$ , the probability that team B wins, assuming that this game does not end up in a tie. Note that for this particular case,

$$P(A) + P(B) = 1 \quad (2.27)$$



## Part II

# Probability

—

|

—

|

—

|

—

|

---

## Part II: Probability

---

*Characterizing Random Variability*

---

*Here we have the opportunity of expounding more clearly what has already been said*

René Descartes (1596–1650)

## Part II: Probability

*Characterizing Random Variability*

- **Chapter 3:** Fundamentals of Probability Theory
- **Chapter 4:** Random Variables
- **Chapter 5:** Multidimensional Random Variables
- **Chapter 6:** Random Variable Transformations
- **Chapter 7:** Application Case Studies I: Probability

# Chapter 3

## Fundamentals of Probability Theory

|       |                                       |    |
|-------|---------------------------------------|----|
| 3.1   | Building Blocks .....                 | 58 |
| 3.2   | Operations .....                      | 60 |
| 3.2.1 | Events, Sets and Set Operations ..... | 61 |
| 3.2.2 | Set Functions .....                   | 64 |
| 3.2.3 | Probability Set Function .....        | 67 |
| 3.2.4 | Final considerations .....            | 68 |
| 3.3   | Probability .....                     | 69 |
| 3.3.1 | The Calculus of Probability .....     | 69 |
| 3.3.2 | Implications .....                    | 71 |
| 3.4   | Conditional Probability .....         | 72 |
| 3.4.1 | Illustrating the Concept .....        | 72 |
| 3.4.2 | Formalizing the Concept .....         | 73 |
| 3.4.3 | Total Probability .....               | 74 |
| 3.4.4 | Bayes' Rule .....                     | 76 |
| 3.5   | Independence .....                    | 77 |
| 3.6   | Summary and Conclusions .....         | 78 |
|       | REVIEW QUESTIONS .....                | 79 |
|       | EXERCISES .....                       | 80 |
|       | APPLICATION PROBLEMS .....            | 84 |

*Before setting out to attack any definite problem  
it behooves us first, without making any selection,  
to assemble those truths that are obvious  
as they present themselves to us  
and afterwards, proceeding step by step,  
to inquire whether any others can be deduced from these.*

René Descartes (1596–1650)

The paradox of randomly varying phenomena — that the aggregate ensemble behavior of unpredictable, irregular, individual observations is stable and regular — provides a basis for developing a systematic analysis approach. Such an approach requires temporarily abandoning the futile task of predicting individual outcomes and instead focussing on characterizing the aggregate ensemble in a mathematically appropriate manner. The central element is a machinery for determining the mathematical probability of the occurrence of each outcome and for quantifying the uncertainty associated with any attempts at predicting the intrinsically unpredictable individual outcomes. How this probability machinery is assembled from a set of simple building blocks and mathematical operations is presented in this chapter, along with the basic concepts required for its subsequent use for systematic analysis of random

phenomena. This chapter is therefore devoted to introducing probability in its basic form *first*, before we begin employing it in subsequent chapters to solve problems involving random phenomena.

---

### 3.1 Building Blocks

A formal mathematical theory for studying random phenomena makes use of certain words, concepts, and terminology in a more restricted technical sense than is typically implied by common usage. We begin by providing the definitions of:

- Experiments; Trials; Outcomes
- Sample space; Events

within the context of the machinery of probability theory .

**1. Experiment:** *Any process that generates observable information about the random phenomenon in question.*

This could be the familiar sort of “experiment” in the sciences and engineering (such as the determination of the pH of a solution, the quantification of the effectiveness of a new drug, or the determination of the effect of an additive on gasoline consumption in an automobile engine); it also includes the simple, almost artificial sort, such as tossing a coin or some dice, drawing a marble from a box, or a card from a well-shuffled deck. We will employ such simple conceptual experiments with some regularity because they are simple and easy to conceive mentally, but more importantly because they serve as useful models for many practical, more complex problems, allowing us to focus on the essentials and avoid getting bogged down with unnecessary and potentially distracting details. For example, in inspecting a manufactured lot for defective parts, so long as the result of interest is whether the selected and tested part is defective or not, the “real experiment” is well-modeled by the toss of an appropriate coin.

**2. Outcome:** *The result of an experiment.*

This could be as simple as an *attribute*, such as the color of a marble drawn from a box, or whether the part drawn from a manufactured lot is defective or not; it could be a *discrete* quantity such as the number of heads observed after 10 tosses of a coin, or the number of contaminants observed on a silicon wafer; it could also be a *continuous* quantity such as the temperature of reactants in a chemical reactor, or the concentration of arsenic in a water sample.

**3. Trial:** A single performance of a well-defined experiment giving rise to an outcome.

Random phenomena are characterized by the fact that each trial of the same experiment performed under identical conditions can potentially produce different outcomes.

Closely associated with the possible outcomes of an experiment and crucial to the development of probability theory are the concepts of the *sample space* and *events*.

**4. Sample Space:** The set of all possible outcomes of an experiment.

If the elements of this set are individual, distinct, countable entities, then the sample space is said to be *discrete*; if, on the other hand, the elements are a continuum of values, the sample space is said to be *continuous*.

**5. Event:** A set of possible outcomes that share a common attribute.

The following examples illustrate these concepts.

**Example 3.1 THE BUILDING BLOCKS OF PROBABILITY**

In tossing a coin 3 times and recording the number of observed heads and tails, identify the experiment, what each trial entails, the outcomes, and the sample space.

**Solution:**

1. Experiment: Toss a coin 3 times; record the number of observed heads (each one as an “H”) and tails (each one as a “T”);
2. Trial: Each trial involves 3 consecutive tosses of the coin;
3. Outcomes: Any one of the following is a possible outcome: HHH, HHT, HTH, THH, HTT, THT, TTH, TTT.
4. Sample space: The set  $\Omega$  defined by

$$\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\} \quad (3.1)$$

consisting of all possible 8 outcomes, is the sample space for this experiment. This is a *discrete* sample space because there are 8 individual, distinct and countable elements.

**Example 3.2 EVENTS ASSOCIATED WITH EXAMPLE 3.1**

Identify some events associated with the experiment introduced in Example 3.1.

**Solution:**

The set  $A = \{HHT, HTH, THH\}$  consists of those outcomes involving the occurrence of *exactly* two heads; it therefore represents the event that “exactly 2 heads are observed” when a coin is tossed 3 times.

The set  $B = \{TTT\}$  consists of the only outcome involving the occurrence of 3 tails; it therefore represents the event that “3 tails are observed”.

The set  $C = \{HHH, HHT, HTH, THH\}$  consists of the outcomes involving the occurrence of *at least* 2 heads; it represents the event that “at least 2 heads are observed”.

Similarly, the set  $D = \{HHH\}$  represents the event that “3 heads are observed”.

A *simple* or *elementary* event is one that consists of one and only one outcome of the experiment; i.e. a set with only one element. Thus, in Example 3.2, set  $B$  and set  $D$  are examples of elementary events. Any other event consisting of more than one outcome is a *complex* or *compound* event. Sets  $A$  and  $C$  in Example 3.2 are compound events. (One must be careful to distinguish between the set and its elements. The set  $B$  in Example 3.2 contains one element, TTT, but the set is not the same as the element. Thus, even though the elementary event consists of a single outcome, one is not the same as the other).

Elementary events possess an important property that is crucial to the development of probability theory:

- An experiment conducted once produces one and only one outcome;
- The elementary event consists of only one outcome;
- One and only one elementary event can occur for every experimental trial;

Therefore:

*Simple (elementary) events are mutually exclusive.*

In Example 3.2, sets  $B$  and  $D$  represent elementary events; observe that if one occurs, the other one cannot. Compound events do not have this property. In this same example, observe that if, after a trial, the outcome is HTH (a tail sandwiched between two heads), event  $A$  has occurred (we have observed precisely 2 heads), but so has event  $C$ , which requires observing 2 or more heads. In the language of sets, the element HTH belongs to both set  $A$  and set  $B$ .

An “elementary event” therefore consists of a single outcome and cannot be decomposed into a simpler event; a “compound event,” on the other hand, consists of a collection of more than one outcome and can therefore be composed from several simple events.

## 3.2 Operations

If rational analysis of random phenomena depends on working with the aggregate ensemble of all possible outcomes, the next step in the assembly of the analytical machinery is a means of operating on the component building blocks identified above. First the outcomes, already represented as events, must be firmly rooted in the mathematical soil of sets so that established basic set operations can be used to operate on events. The same manipulations of standard algebra and the algebra of sets can then be used to obtain algebraic relationships between the events that comprise the aggregate ensemble of the random phenomenon in question. The final step is the definition of functions and accompanying operational rules that allow us to perform functional analysis on the events.

### 3.2.1 Events, Sets and Set Operations

We earlier defined the sample space  $\Omega$  as a set whose elements are all the possible outcomes of an experiment. Events are also sets, but they consist of only certain elements from  $\Omega$  that share a common attribute. Thus,

*Events are subsets of the sample space.*

Of all the subsets of  $\Omega$ , there are two special ones with important connotations:  $\Phi$ , the empty set consisting of no elements at all, and  $\Omega$  itself. In the language of events, the former represents the *impossible* event, while the latter represents the *certain* event.

Since they are sets, events are amenable to analysis using precisely the same algebra of set operations — union, intersection and complement — which we now briefly review.

**1. Union:**  $A \cup B$  represents the set of elements that are either in  $A$  or  $B$ . In general,

$$A_1 \cup A_2 \cup A_3 \dots \cup A_k = \bigcup_{i=1}^k A_i \quad (3.2)$$

is the set of elements that are in *at least* one of the  $k$  sets,  $\{A_i\}_1^k$ .

**2. Intersection:**  $A \cap B$  represents the set of elements that are in *both*  $A$  and

B. In general,

$$A_1 \cap A_2 \cap A_3 \dots \cap A_k = \bigcap_{i=1}^k A_i \quad (3.3)$$

is the set of elements that are *common* to all the  $k$  sets,  $\{A_i\}_1^k$ .

To discuss the third set operation requires two special sets: The universal set (or “universe”), typically designated  $\Omega$ , and the null (or empty) set, typically designated  $\Phi$ . The universal set consists of all possible elements of interest, while the null set contains no elements. (We have just recently introduced such sets above but in the specific context of the sample space of an experiment; the current discussion is general and not restricted to the analysis of randomly varying phenomena and their associated sample spaces.)

These sets have the special properties that for any set  $A$ ,

$$A \cup \Phi = A; \quad (3.4)$$

$$A \cap \Phi = \Phi \quad (3.5)$$

and

$$A \cup \Omega = \Omega; \quad (3.6)$$

$$A \cap \Omega = A \quad (3.7)$$

**3. Complement:**  $A^*$ , the complement of set  $A$ , is always defined with respect to the universal set  $\Omega$ ; it consists of all the elements of  $\Omega$  that are *not* in  $A$ . The following are basic relationships associated with the complement operation:

$$\Omega^* = \Phi; \Phi^* = \Omega \quad (3.8)$$

$$(A^*)^* = A; \quad (3.9)$$

$$(A \cup B)^* = A^* \cap B^* \quad (3.10)$$

$$(A \cap B)^* = A^* \cup B^* \quad (3.11)$$

with the last two expressions known as DeMorgan’s Laws.

The rules of set algebra (similar to those of standard algebra) are as follows:

- **Commutative law:**

$$A \cup B = B \cup A$$

$$A \cap B = B \cap A$$

- **Associative law:**

$$(A \cup B) \cup C = A \cup (B \cup C)$$

$$(A \cap B) \cap C = A \cap (B \cap C)$$

- **Distributive Law:**

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$$

$$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$$

**TABLE 3.1:** Subsets and Events

| Subset     | Event                     |
|------------|---------------------------|
| $\Omega$   | Certain event             |
| $\Phi$     | Impossible event          |
| $A^*$      | Non-occurrence of event A |
| $A \cup B$ | Event A or B              |
| $A \cap B$ | Events A and B            |

The following table presents some information about the nature of subsets of  $\Omega$  interpreted in the language of events.

Note in particular that if  $A \cap B = \Phi$ , A and B are said to be disjoint sets (with no elements in common); in the language of events, this implies that event A occurring together with event B is *impossible*. Under these circumstances, events A and B are said to be mutually exclusive.

### Example 3.3 PRACTICAL ILLUSTRATION OF SETS AND EVENTS

Samples from various batches of a polymer resin manufactured at a plant site are tested in a quality control laboratory before release for sale. The result of the tests allows the manufacturer to classify the product into the following 3 categories:

1. Meets or exceeds quality requirement; Assign #1; approve for sale as 1<sup>st</sup> quality.
2. Barely misses quality requirement; Assign #2; approve for sale as 2<sup>nd</sup> grade at a lower price.
3. Fails completely to meet quality requirement; Assign #3; reject as poor grade and send back to be incinerated.

Identify the experiment, outcome, trial, sample space and the events associated with this practical problem.

#### Solution:

1. Experiment: Take a sample of polymer resin and carry out the prescribed product quality test.
2. Trial: Each trial involves taking a *representative* sample from each polymer resin batch and testing it as prescribed.
3. Outcomes: The assignment of a number 1, 2, or 3 depending on how the result of the test compares to the product quality requirements.
4. Sample space: The set  $\Omega = \{1, 2, 3\}$  containing all possible outcomes.
5. Events: The subsets of the sample space are identified as follows:  $E_0 = \{\Phi\}$ ;  $E_1 = \{1\}$ ;  $E_2 = \{2\}$ ;  $E_3 = \{3\}$ ;  $E_4 = \{1, 2\}$ ;  $E_5 = \{1, 3\}$ ;  $E_6 = \{2, 3\}$ ;  $E_7 = \{1, 2, 3\}$ . Note that there are 8 in all. In general, a set with  $n$  distinct elements will have  $2^n$  subsets.

Note that this “real” experiment is identical in spirit to the “conceptual” experiment in which 3 identical ping-pong balls inscribed with the numbers 1, 2, and 3 are placed in a box, and each trial involves drawing one out and recording the inscribed number found on the chosen ball. Employing the “artificial” surrogate may sometimes be a useful device to enable us focus on the essential components of the problem.

**Example 3.4 INTERPRETING EVENTS OF EXAMPLE 3.3**

Provide a practical interpretation of the events identified in the quality assurance problem of Example 3.3 above.

**Solution:**

$E_1 = \{1\}$  is the event that the batch is of 1<sup>st</sup> grade;

$E_2 = \{2\}$  is the event that the batch is of 2<sup>nd</sup> grade;

$E_3 = \{3\}$  is the event that the batch is rejected as poor grade.

These are elementary events; they are mutually exclusive.

$E_4 = \{1, 2\}$  is the event that the batch is either 1<sup>st</sup> grade or 2<sup>nd</sup> grade;

$E_5 = \{1, 3\}$  is the event that the batch is either 1<sup>st</sup> grade or rejected;

$E_6 = \{2, 3\}$  is the event that the batch is either 2<sup>nd</sup> grade or rejected.

These events are not elementary and are not mutually exclusive. For instance, if a sample analysis indicates the batch is 1<sup>st</sup> grade, then the events  $E_1, E_4$  and  $E_5$  have all occurred.

$E_7 = \{1, 2, 3\} = \Omega$  is the event that the batch is either 1<sup>st</sup> grade or 2<sup>nd</sup> grade, or rejected;

$E_0 = \Phi$  is the event that the batch is neither 1<sup>st</sup> grade nor 2<sup>nd</sup> grade, nor rejected.

Event  $E_7$  is certain to happen: the outcome of the experiment has to be one of these three classifications—there is no other alternative; event  $E_0$  on the other hand is impossible, for the same reason.

**Example 3.5 COMPOUND EVENTS FROM ELEMENTARY EVENTS**

Show how the compound events in Examples 3.3 and 3.4 can be composed from (or decomposed into) elementary events.

**Solution:**

The compound events  $E_4, E_5, E_6$  and  $E_7$  are related to the elementary events  $E_1, E_2$  and  $E_3$  as follows:

$$E_4 = E_1 \cup E_2 \quad (3.12)$$

$$E_5 = E_1 \cup E_3 \quad (3.13)$$

$$E_6 = E_2 \cup E_3 \quad (3.14)$$

$$E_7 = E_1 \cup E_2 \cup E_3 \quad (3.15)$$

**TABLE 3.2:** Class list and attributes

| Name    | Sex<br>(M or F) | Age<br>(in years) | Amount in wallet<br>(to the nearest \$) | Height<br>(in inches) |
|---------|-----------------|-------------------|---|-----------------------|
| Allison | F               | 21                | \$ 17.00                                | 66                    |
| Ben     | M               | 23                | \$ 15.00                                | 72                    |
| Chrissy | F               | 23                | \$ 26.00                                | 65                    |
| Daoud   | M               | 25                | \$ 35.00                                | 67                    |
| Evan    | M               | 22                | \$ 27.00                                | 73                    |
| Fouad   | M               | 20                | \$ 15.00                                | 69                    |
| Gopalan | M               | 21                | \$ 29.00                                | 68                    |
| Helmut  | M               | 19                | \$ 13.00                                | 71                    |
| Ioannis | M               | 25                | \$ 32.00                                | 70                    |
| Jim     | M               | 24                | \$ 53.00                                | 74                    |
| Katie   | F               | 22                | \$ 41.00                                | 70                    |
| Larry   | M               | 24                | \$ 28.00                                | 72                    |
| Moe     | M               | 21                | \$ 18.00                                | 71                    |
| Nathan  | M               | 22                | \$ 6.00                                 | 68                    |
| Olu     | M               | 26                | \$ 23.00                                | 72                    |

### 3.2.2 Set Functions

A function  $F(\cdot)$ , defined on the subsets of  $\Omega$  such that it assigns one and only one real number to each subset of  $\Omega$ , is known as a *set function*. By this definition, no one subset can be assigned more than one number by a set function. The following examples illustrate the concept.

#### Example 3.6 SET FUNCTIONS DEFINED ON THE SET OF STUDENTS IN A CLASSROOM

The following table shows a list of attributes associated with 15 students in attendance on a particular day in a 600 level course offered at the University of Delaware. Let set  $A$  be the subset of female students and  $B$ , the subset of male students. Obtain the real number assigned by the following set functions:

1.  $N(A)$ , the total number of female students in class;
2.  $N(\Omega)$ , the total number of students in class;
3.  $M(B)$ , the sum total amount of money carried by the male students;
4.  $\bar{H}(A)$ , the average height (in inches) of female students;
5.  $Y^+(B)$ , the maximum age, in years, of male students

#### Solution:

1.  $N(A) = 3$ ;
2.  $N(\Omega) = 15$ ;
3.  $M(B) = \$293.00$ ;
4.  $\bar{H}(A) = 67$  ins.

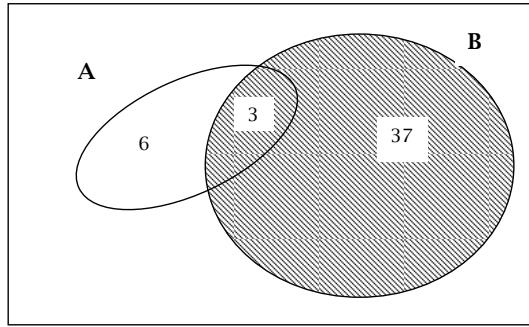


FIGURE 3.1: Venn Diagram for Example 3.7

$$5. Y^+(B) = 26 \text{ years.}$$

A set function  $Q$  is said to be *additive* if for every pair of disjoint subsets  $A$  and  $B$  of  $\Omega$ ,

$$Q(A \cup B) = Q(A) + Q(B) \quad (3.16)$$

For example, the set function  $N(\cdot)$  in Example 3.6 is an additive set function. Observe that the sets  $A$  and  $B$  in this example are disjoint; furthermore  $\Omega = A \cup B$ . Now,  $N(\Omega) = N(A \cup B) = 15$  while  $N(A) = 3$  and  $N(B) = 12$ . Thus for this example,

$$N(A \cup B) = N(A) + N(B) \quad (3.17)$$

However,  $\bar{H}(\cdot)$  is *not* an additive set function, and neither is  $Y^+(\cdot)$ .

In general, when two sets are not disjoint, i.e. when  $A \cap B \neq \emptyset$ , so that the intersection is non-empty, it is easy to show (see exercise at the end of the chapter) that if  $Q(\cdot)$  is an additive set function,

$$Q(A \cup B) = Q(A) + Q(B) - Q(A \cap B) \quad (3.18)$$

### Example 3.7 ADDITIVE SET FUNCTION ON NON-DISJOINT SETS

An old batch of spare parts contains 40 parts, of which 3 are defective; a newly manufactured batch of 60 parts was added to make up a consolidated batch of 100 parts, of which a total of 9 are defective. Find the total number of parts that are either defective or from the old batch.

#### Solution:

If  $A$  is the set of defective parts and  $B$  is the set of parts from the old batch, and if  $N(\cdot)$  is the number of parts in a set, then we seek  $N(A \cup B)$ . The Venn diagram in Fig 3.1 shows the distribution of elements in each set.

From Eq (3.18),

$$N(A \cup B) = N(A) + N(B) - N(A \cap B) = 9 + 40 - 3 = 46 \quad (3.19)$$

so that there are 46 parts that are either defective or from the old batch.

### 3.2.3 Probability Set Function

Let  $P(\cdot)$  be an additive set function defined on all subsets of  $\Omega$ , the sample space of all the possible outcomes of an experiment, such that:

1.  $P(A) \geq 0$  for every  $A \subset \Omega$ ;
2.  $P(\Omega) = 1$ ;
3.  $P(A \cup B) = P(A) + P(B)$  for all mutually exclusive events  $A$  and  $B$

then  $P(\cdot)$  is a *probability set function*.

Remarkably, these three simple rules (axioms) due to Kolmogorov, are sufficient to develop the mathematical theory of probability. The following are important properties of  $P(\cdot)$  arising from these axioms.

1. To each event  $A$ , it assigns a non-negative number,  $P(A)$ , its probability;
2. To the certain event  $\Omega$ , it assigns unit probability;
3. The probability that either one or the other of two mutually exclusive events  $A, B$  will occur is the sum of the probabilities that each event will occur.

The following corollaries are important consequences of the foregoing three axioms:

**Corollary 1.**  $P(A^*) = 1 - P(A)$ .

The probability of non-occurrence of  $A$  is 1 minus the probability of its occurrence. Equivalently, the combined probability of the occurrence of an event and of its non-occurrence add up to 1. This follows from the fact that

$$\Omega = A \cup A^*; \quad (3.20)$$

that  $A$  and  $A^*$  are disjoint sets; that  $P(\cdot)$  is an additive set function, and that  $P(\Omega) = 1$ .

**Corollary 2.**  $P(\Phi) = 0$ .

The probability of an impossible event occurring is zero. This follows from the fact that  $\Omega^* = \Phi$  and from corollary 1 above.

**Corollary 3.**  $A \subset B \Rightarrow P(A) \leq P(B)$ .

If  $A$  is a subset of  $B$  then the probability of occurrence of  $A$  is less than, or equal to, the probability of the occurrence of  $B$ . This follows from the fact that under these conditions,  $B$  can be represented as the union of 2 disjoint sets:

$$B = A \cup (B \cap A^*) \quad (3.21)$$

and from the additivity of  $P(.)$ ,

$$P(B) = P(A) + P(B \cap A^*) \quad (3.22)$$

so that from the non-negativity of  $P(.)$ , we obtain,

$$P(B) \geq P(A) \quad (3.23)$$

**Corollary 4.**  $0 \leq P(A) \leq 1$  for all  $A \subset \Omega$ .

The probability of any realistic event occurring is bounded between zero and 1. This follows directly from the first 2 axioms and from corollary 3 above.

**Corollary 5.**  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$  for any pair of subsets  $A$  and  $B$ .

This follows directly from the additivity of  $P(.)$  and results presented earlier in Eq (3.18).

### 3.2.4 Final considerations

Thus far, in assembling the machinery for dealing with random phenomena by characterizing the aggregate ensemble of all possible outcomes, we have encountered the sample space  $\Omega$ , whose elements are all the possible outcomes of an experiment; we have presented events as collections of these outcomes (and hence subsets of  $\Omega$ ); and finally  $P(.)$ , the probability set function defined on subsets of  $\Omega$ , allows the axiomatic definition of the probability of an event.

What we need next is a method for actually obtaining any particular probability  $P(A)$  once the event  $A$  has been defined. Before we can do this, however, for completeness, a set of final considerations are in order.

Even though as presented, events are subsets of  $\Omega$ , not all subsets of  $\Omega$  are events. There are all sorts of subtle mathematical reasons for this, including the (somewhat unsettling) case in which  $\Omega$  consists of infinitely many elements, as is the case when the outcome is a continuous entity and can therefore take on values on the real line. In this case, clearly,  $\Omega$  is the set of all real numbers. A careful treatment of these issues requires the introduction of Borel fields (see for example, Kingman and Taylor, 1966, Chapter 11<sup>1</sup>). This is necessary because, as the reader may have anticipated, the calculus of probability requires making use of set operations, unions and intersections, as well as sequences and limits of events. As a result, it is important that sets resulting from such operations are themselves events. This is strictly true of Borel fields.

Nevertheless, for all practical purposes, and most practical applications, it is often not necessary to distinguish between the subsets of  $\Omega$  and “genuine” events. For the reader willing to accept on faith the end result—the probability

---

<sup>1</sup>Kingman, J.F.C. and Taylor, S.J., *Introduction to the Theory of Measure and Probability*, Cambridge University Press, 1966.

distribution function presented fully in Chapters 4 and 5—a lack of detailed knowledge of such subtle, but important, fine points will not constitute a hinderance to the appropriate use of the tool.

### 3.3 Probability

We are now in a position to discuss how to use the machinery we have assembled above to determine the probability of any particular event  $A$ .

#### 3.3.1 The Calculus of Probability

Once the sample space  $\Omega$  for any random experiment has been specified and the events (subsets of the sample space) identified, the following is the procedure for determining the probability of any event  $A$ , based on the important property that elementary events are *mutually exclusive*:

- Assign probabilities to all the elementary events in  $\Omega$ ;
- Determine the probability of any compound event from the probability of the elementary events making up the compound event of interest.

The procedure is particularly straightforward to illustrate for discrete sample spaces with a countable number of elements. For example, if  $\Omega = \{d_1, d_2, \dots, d_N\}$  consists of  $N$  outcomes, then there are  $N$  elementary events,  $E_i = \{d_i\}$ . To each of these elementary events, we assign the probability  $p_i$  (we will discuss shortly *how* such assignments are made) subject to the constraint that  $\sum_i^N p_i = 1$ . From here, if

$$A = \{d_1, d_2, d_4\} \quad (3.24)$$

and if  $P(A)$  represents the probability of event  $A$  occurring, then,

$$P(A) = p_1 + p_2 + p_4 \quad (3.25)$$

and for

$$B = \{d_3, d_5, \dots, d_N\} \quad (3.26)$$

then

$$P(B) = 1 - p_1 - p_2 - p_4 \quad (3.27)$$

The following examples illustrate how probabilities  $p_i$  may be assigned to elementary events.

**Example 3.8 ASSIGNMENTS FOR EQUIPROBABLE OUTCOMES**

The experiment of tossing a coin 3 times and recording the observed number of heads and tails was considered in Examples 3.1 and 3.2. There the sample space was obtained in Eq (4.5) as:

$$\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}, \quad (3.28)$$

a set with 8 elements that comprise all the possible outcomes of the experiment. Several events associated with this experiment were identified in Example 3.2.

If there is no reason for any one of the 8 possible outcomes to be any more likely to occur than any other one, the outcomes are said to be *equiprobable* and we assign a probability of  $1/8$  to each one. This gives rise to the following equiprobable assignment of probability to the 8 elementary events:

$$\begin{aligned} P(E_1) &= P\{HHH\} = 1/8 \\ P(E_2) &= P\{HHT\} = 1/8 \\ P(E_3) &= P\{HTH\} = 1/8 \\ &\vdots \\ P(E_7) &= P\{TTH\} = 1/8 \\ P(E_8) &= P\{TTT\} = 1/8 \end{aligned} \quad (3.29)$$

Note that

$$\sum_1^8 p_i = \sum_1^8 P(E_i) = 1 \quad (3.30)$$

And now because the event

$$A = \{HHT, HTH, THH\} \quad (3.31)$$

identified in Example 3.2 (the event that exactly 2 heads are observed) consists of three elementary events  $E_2, E_3$  and  $E_4$ , so that

$$A = E_2 \cup E_3 \cup E_4, \quad (3.32)$$

because these sets are disjoint, we have that

$$P(A) = P(E_2) + P(E_3) + P(E_4) = 3/8 \quad (3.33)$$

Similarly, for the events  $B, C$  and  $D$  identified in Example 3.2, we have, respectively,  $P(C) = 4/8 = 0.5$  and  $P(B) = P(D) = 1/8$ .

Other means of probability assignment are possible, as illustrated by the following example.

**Example 3.9 ALTERNATIVE ASSIGNMENTS FROM A-PRIORI KNOWLEDGE**

Consider the manufacturing example discussed in Examples 3.3 and 3.4.

Suppose that historically 75% of manufactured batches have been of 1<sup>st</sup> grade, 15% of grade 2 and the rest rejected. Assuming that nothing has changed in the manufacturing process, use this information to assign probabilities to the elementary events identified in Example 3.3, and determine the probabilities for all the possible events associated with this problem.

**Solution:**

Recall from Examples 3.3 and 3.4 that the sample space in this case is  $\Omega = \{1, 2, 3\}$  containing all 3 possible outcomes; the 3 elementary events are  $E_1 = \{1\}$ ;  $E_2 = \{2\}$ ;  $E_3 = \{3\}$ ; the other events (the remaining subsets of the sample space) had been previously identified as:  $E_0 = \{\Phi\}$ ;  $E_4 = \{1, 2\}$ ;  $E_5 = \{1, 3\}$ ;  $E_6 = \{2, 3\}$ ;  $E_7 = \{1, 2, 3\}$ .

From the provided information, observe that it is entirely reasonable to assign probabilities to the elementary events as follows:

$$P(E_1) = 0.75 \quad (3.34)$$

$$P(E_2) = 0.15 \quad (3.35)$$

$$P(E_3) = 0.10 \quad (3.36)$$

Note that these probabilities sum to 1 as required. From here we may now compute the probabilities for the other events:

$$P(E_4) = P(E_1) + P(E_2) = 0.9 \quad (3.37)$$

$$P(E_5) = P(E_1) + P(E_3) = 0.85 \quad (3.38)$$

$$P(E_6) = P(E_2) + P(E_3) = 0.25 \quad (3.39)$$

For completeness, we note that  $P(E_0) = 0$ ; and  $P(E_7) = 1$ .

### 3.3.2 Implications

It is worth spending a few moments to reflect on the results obtained from this last example.

The premise is that the manufacturing process is subject to many sources of variability so that despite having an objective of maintaining consistent product quality, its product may still fall into any one of the three quality grade levels in an unpredictable manner. Nevertheless, even though the particular grade (outcome) of any particular tested sample (experiment) is uncertain and unpredictable, this example shows us how we can determine the probability of the occurrence of the entire collection of all possible events. First, the more obvious elementary events: for example the probability that a sample will be grade 1 is 0.75. Even the less obvious complex events have also been characterized. For example, if we are interested in the probability of making any money at all on what is currently being manufactured, this is the event  $E_4$  (producing saleable grade 1 or 2 material); the answer is 0.9. The probability

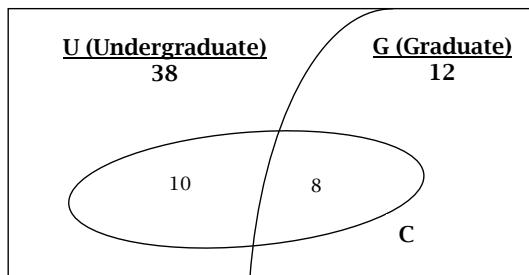


FIGURE 3.2: Venn diagram of students in a thermodynamics class

of *not* making grade 1 material is 0.25 (the non-occurrence of event  $E_1$ , or equivalently the event  $E_6$ ).

With this example, what we have actually done is to construct a “model” of how the probability of the occurrence of events is “distributed” over the entire collection of all possible events. In subsequent chapters, we make extensive use of the mechanism illustrated here in developing probability models for complex random phenomena, proceeding from the probability of elementary events and employing the calculus of probability to obtain the required “probability distribution” expressions.

### 3.4 Conditional Probability

#### 3.4.1 Illustrating the Concept

Consider a chemical engineering thermodynamics class consisting of 50 total students of which 38 are undergraduates and the rest are graduate students. Of the 12 graduate students, 8 are chemistry students; of the 38 undergraduates, 10 are chemistry students. We may define the following sets:

- $\Omega$ , the (universal) set of all students (50 elements);
- $G$ , the set of graduate students (12 elements);
- $C$ , the set of chemistry students (18 elements)

Note that the set  $G \cap C$ , the set of graduate chemistry students, contains 8 elements. (See Fig 3.2.)

We are interested in the following problem: select a student at random; given that the choice results in a chemistry student, what is the probability that she/he is a graduate student? This is a problem of finding the probability of the occurrence of an event *conditioned* upon the prior occurrence of another one.

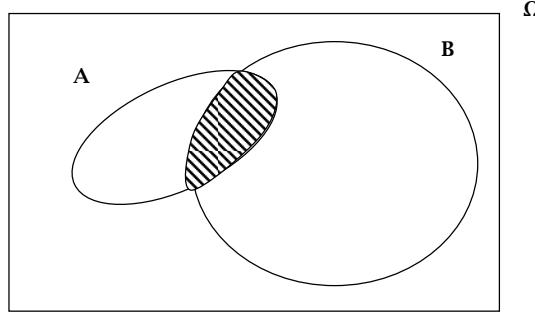


FIGURE 3.3: The role of “conditioning” Set  $B$  in conditional probability

In this particular case, the total number of students in the chemistry group is 18, of which 8 are graduates. The required probability is thus precisely that of choosing one of the 8 graduate students out of all the possible 18 chemistry students; and, assuming equiprobable outcomes, this probability is  $8/18$ . (Note also from the definition of the sets above, that  $P(C) = 18/50$  and  $P(G \cap C) = 8/50$ .)

We may now formalize the just illustrated concept as follows.

### 3.4.2 Formalizing the Concept

For two sets  $A$  and  $B$ , the conditional probability of  $A$  given  $B$ , denoted  $P(A|B)$ , is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (3.40)$$

where  $P(B) > 0$ .

Observe how the set  $B$  now plays the role that  $\Omega$  played in “unconditional” probability (See Fig 3.3); in other words, the process of “conditioning” restricts the set of relevant outcomes to  $B \subset \Omega$ . In this sense,  $P(A)$  is really  $P(A|\Omega)$ , which, according to Eq. (5.33), may be written as

$$P(A|\Omega) = \frac{P(A \cap \Omega)}{P(\Omega)} = \frac{P(A)}{1} \quad (3.41)$$

Returning now to the previous illustration, we see that the required quantity is  $P(G|C)$ , and by definition,

$$P(G|C) = \frac{P(G \cap C)}{P(C)} = \frac{8/50}{18/50} = 8/18 \quad (3.42)$$

as obtained previously. The unconditional probability  $P(G)$  is  $12/50$ .

The conditional probability  $P(A|B)$  possesses all the required properties of a probability set function defined on subsets of  $B$ :

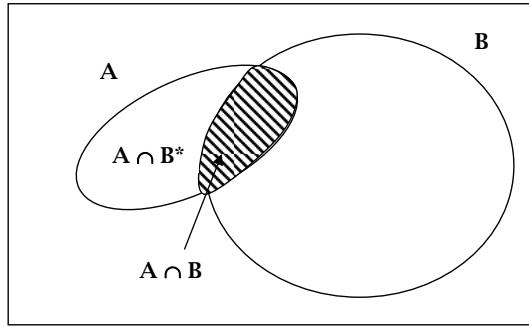


FIGURE 3.4: Representing set A as a union of 2 disjoint sets

1.  $0 < P(A|B) \leq 1$ ;
2.  $P(B|B) = 1$ ;
3.  $P(A_1 \cup A_2|B) = P(A_1|B) + P(A_2|B)$  for disjoint  $A_1$  and  $A_2$ .

The following identities are easily derived from the definition given above for  $P(A|B)$ :

$$P(A \cap B) = P(B)P(A|B); P(B) > 0 \quad (3.43)$$

$$= P(A)P(B|A); P(A) > 0 \quad (3.44)$$

Conditional probability is a particularly important concept in science and engineering applications because we often have available to us some *a-priori* knowledge about a phenomenon; the required probabilities then become conditioned upon the available information.

### 3.4.3 Total Probability

It is possible to obtain total probabilities when only conditional probabilities are available. We now present some very important results relating conditional probabilities to total probability.

Consider events  $A$  and  $B$ , not necessarily disjoint. From the Venn diagram in Fig 3.4, we may write  $A$  as the union of 2 disjoint sets as follows:

$$A = (A \cap B) \cup (A \cap B^*) \quad (3.45)$$

In words, this expression states that the points in  $A$  are made up of two groups: the points in  $A$  that are also in  $B$ , and the points in  $A$  that are not in  $B$ . And because the two sets are disjoint, so that the events they represent are mutually exclusive, we have:

$$P(A) = P(A \cap B) + P(A \cap B^*) \quad (3.46)$$

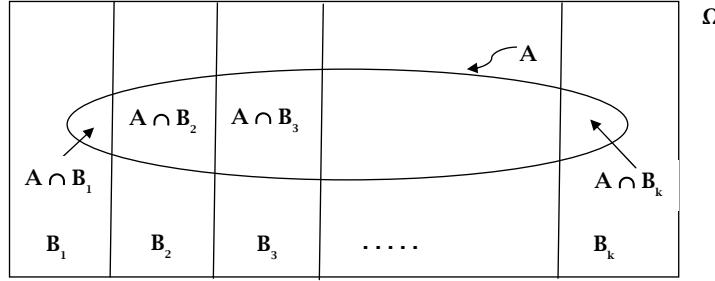


FIGURE 3.5: Partitioned sets for generalizing total probability result

and from the definition of conditional probability, we obtain:

$$P(A) = P(A|B)P(B) + P(A|B^*)P(B^*) \quad (3.47)$$

or, alternatively,

$$P(A) = P(A|B)P(B) + P(A|B^*)[1 - P(B)] \quad (3.48)$$

This powerful result states that the (unconditional, or total) probability of an event  $A$  is a weighted average of two partial (or conditional) probabilities: the probability conditioned on the occurrence of  $B$  and the probability conditioned upon the non-occurrence of  $B$ ; the weights, naturally, are the respective probabilities of the “conditioning” event.

This may be generalized as follows: First we partition  $\Omega$  into a union of  $k$  disjoint sets:

$$\Omega = B_1 \cup B_2 \cup B_3 \cup \dots \cup B_k = \bigcup_{i=1}^k B_i \quad (3.49)$$

For any  $A$  that is an arbitrary subset of  $\Omega$ , observe that

$$A = (A \cap B_1) \cup (A \cap B_2) \cup (A \cap B_3) \cup \dots \cup (A \cap B_k) \quad (3.50)$$

which is a partitioning of the set  $A$  as a union of  $k$  disjoint sets (See Fig 3.5). As a result,

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \dots + P(A \cap B_k) \quad (3.51)$$

but since

$$P(A \cap B_i) = P(A|B_i)P(B_i) \quad (3.52)$$

we immediately obtain

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \dots + P(A|B_k)P(B_k) \quad (3.53)$$

Thus:

$$P(A) = \sum_{i=1}^k P(A|B_i)P(B_i) \quad (3.54)$$

an expression that is sometimes referred to as the “Theorem of total probability” used to compute total probability  $P(A)$  from  $P(A|B_i)$  and  $P(B_i)$ .

The following example provides an illustration.

**Example 3.10 TOTAL PROBABILITY**

A company manufactures light bulbs of 3 different types ( $T_1, T_2, T_3$ ) some of which are defective right from the factory. From experience with the manufacturing process, it is known that the fraction of defective Type 1 bulbs is 0.1; Types 2 and 3 have respective defective fractions of  $1/15$  and 0.2.

A batch of 200 bulbs were sent to a quality control laboratory for testing: 100 Type 1, 75 Type 2, and 25 Type 3. What is the probability of finding a defective bulb?

**Solution:**

The supplied information may be summarized as follows: Prior conditional probabilities of defectiveness,

$$P(D|T_1) = 0.1; P(D|T_2) = 1/15; P(D|T_3) = 0.2; \quad (3.55)$$

and the distribution of numbers of bulb types in the test batch:

$$N(T_1) = 100; N(T_2) = 75; N(T_3) = 25. \quad (3.56)$$

Assuming equiprobable outcomes, this number distribution immediately implies the following:

$$P(T_1) = 100/200 = 0.5; P(T_2) = 0.375; P(T_3) = 0.125 \quad (3.57)$$

From the expression for total probability in Eq.(3.53), we have:

$$P(D) = P(D|T_1)P(T_1) + P(D|T_2)P(T_2) + P(D|T_3)P(T_3) = 0.1 \quad (3.58)$$

### 3.4.4 Bayes' Rule

A question of practical importance in many applications is:

*Given  $P(A|B_i)$  and  $P(B_i)$ , how can we obtain  $P(B_i|A)$ ?*

In other words, how can we “reverse” probabilities?

The total probability expression we have just derived provides a way to answer this question. Note from the definition of conditional probability that:

$$P(B_i|A) = \frac{P(B_i \cap A)}{P(A)} \quad (3.59)$$

but

$$P(B_i \cap A) = P(A \cap B_i) = P(A|B_i)P(B_i) \quad (3.60)$$

which, when substituted into (3.59), gives rise to a very important result:

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^k P(A|B_i)P(B_i)} \quad (3.61)$$

This famous result, due to the Revd. Thomas Bayes (1763), is known as “Bayes’ Rule” and we will encounter it again in subsequent chapters. For now, it is an expression that can be used to compute the (unknown) *à-posteriori* probability  $P(B_i|A)$  of events  $B_i$  from the *à-priori* probabilities  $P(B_i)$  and the (known) conditional probabilities  $P(A|B_i)$ . It indicates that the unknown *à-posteriori* probability is proportional to the product of the *à-priori* probability and the known conditional probability we wish to reverse; the constant of proportionality is the reciprocal of the total probability of event  $A$ .

This result is the basis of an alternative approach to data analysis (discussed in Section 14.6 of Chapter 14) wherein available prior information is incorporated in a systematic fashion into the analysis of experimental data.

### 3.5 Independence

For two events  $A$  and  $B$ , the conditional probability  $P(A|B)$  was defined earlier in Eq.(5.33). In general, this conditional probability will be different from the unconditional probability  $P(A)$ , indicating that the knowledge that  $B$  has occurred affects the probability of the occurrence of  $A$ .

However, when the occurrence of  $B$  has no effect on the occurrence of  $A$ , then the events  $A$  and  $B$  are said to be independent and

$$P(A|B) = P(A) \quad (3.62)$$

so that the conditional and unconditional probabilities are identical. This will occur when

$$\frac{P(A \cap B)}{P(B)} = P(A) \quad (3.63)$$

so that

$$P(A \cap B) = P(A)P(B) \quad (3.64)$$

Thus, when events  $A$  and  $B$  are independent, the probability of the two events happening concurrently is the product of the probabilities of each one occurring by itself. Note that the expression in Eq.(3.64) is symmetric in  $A$  and  $B$  so that if  $A$  is independent of  $B$ , then  $B$  is also independent of  $A$ .

This is another in the collection of very important results used in the

development of probability models. We already encountered the first one: that when two events  $A$  and  $B$  are *mutually exclusive*,  $P(A \text{ or } B) = P(A \cup B) = P(A) + P(B)$ . Under these circumstance,  $P(A \cap B) = 0$ , since the event  $A$  occurring together with event  $B$  is impossible when  $A$  and  $B$  are mutually exclusive. Eq.(3.64) is the complementary result: that when two events are *independent*  $P(A \text{ and } B) = P(A \cap B) = P(A)P(B)$ .

Extended to three events, the result states that the events  $A, B, C$  are independent if all of the following conditions hold:

$$P(A \cap B) = P(A)P(B) \quad (3.65)$$

$$P(B \cap C) = P(B)P(C) \quad (3.66)$$

$$P(A \cap C) = P(A)P(C) \quad (3.67)$$

$$P(A \cap B \cap C) = P(A)P(B)P(C) \quad (3.68)$$

implying more than just pairwise independence.

### 3.6 Summary and Conclusions

This chapter has been primarily concerned with assembling the machinery of probability from the building blocks of events in the sample space,  $\Omega$ —the collection of all possible randomly varying outcomes of an experiment. We have seen how the probability of an event  $A$  arises naturally from the probability set function, an additive set function defined on the set  $\Omega$  that satisfies the three axioms of Kolmogorov.

Having established the concept of probability and how the probability of any subset of  $\Omega$  can be computed, a straightforward extension to special events restricted to “conditioning sets” in  $\Omega$  led to the related concept of conditional probability. The idea of total probability, the result known as Bayes’ rule, and especially the concept of independence all arise naturally from conditional probability and have profound consequences for random phenomena analysis that cannot be fully appreciated until much later.

We note in closing that the presentation of probability in this chapter (especially as a tool for solving problems involving randomly varying phenomena) is still quite rudimentary because the development is not quite complete yet. The final step in the development of the probability machinery, undertaken primarily in the next chapter, requires the introduction of the random variable,  $X$ , from which the analysis tool, the probability distribution function,  $f(x)$ , emerges and is fully characterized.

Here are some of the main points of the chapter again:

- Events, as subsets of the sample space,  $\Omega$ , can be elementary (simple) or compound (complex); if elementary, then they are mutually exclusive; if compound, then they can be composed from several simple events.

- Once probabilities have been assigned to *all* elementary events in  $\Omega$ , then  $P(A)$ , the probability of any other subset  $A$  of  $\Omega$ , can be determined on the basis of the probability set function  $P(.)$  defined on all subsets of  $\Omega$  according to the three axioms of Kolmogorov:
  1.  $P(A) \geq 0$  for every  $A \subset \Omega$ ;
  2.  $P(\Omega) = 1$ ;
  3.  $P(A \cup B) = P(A) + P(B)$  for all mutually exclusive events  $A$  and  $B$ .
- Conditional Probability:* For any two events  $A$  and  $B$  in  $\Omega$ , the conditional probability  $P(A|B)$  is given by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- Total Probability:* Given conditional (partial) probabilities  $P(A|B_i)$ , and  $P(B_i)$  for each conditioning set, the unconditional (total) probability of  $A$  is given by

$$P(A) = \sum_{i=1} P(A|B_i)P(B_i)$$

- Mutual Exclusivity:* Two events  $A$  and  $B$  are mutually exclusive if

$$P(A \cup B) = P(A) + P(B),$$

in which case  $P(A \cap B) = 0$ .

- Independence:* Two events  $A$  and  $B$  are independent if

$$P(A \cap B) = P(A)P(B)$$

## REVIEW QUESTIONS

1. What are the five basic building blocks of probability theory as presented in Section 3.1? Define each one.
2. What is a simple (or elementary) event and how is it different from a complex (or compound) event?
3. Why are elementary events mutually exclusive?
4. What is the relationship between events and the sample space?
5. In the language of events, what does the empty set,  $\Phi$ , represent? What does the entire sample space,  $\Omega$ , represent?

- 6.** Given two sets  $A$  and  $B$ , in the language of events, what do the following sets represent:  $A^*$ ;  $A \cup B$ ; and  $A \cap B$ ?
- 7.** What does it mean that two events  $A$  and  $B$  are mutually exclusive?
- 8.** What is a set function in general and what is an *additive* set function in particular?
- 9.** What are the three fundamental properties of a probability set function (also known as Kolmogorov's axioms)?
- 10.** How is the probability of any event  $A \subset \Omega$  determined from the elements events in  $\Omega$ ?
- 11.** For any two sets  $A$  and  $B$ , what is the definition of  $P(A|B)$ , the conditional probability of  $A$  given  $B$ ? If the two sets are disjoint such that  $A \cap B = \emptyset$ , in words, what does  $P(A|B)$  mean in this case?
- 12.** How does one obtain total probability from partial (i.e., conditional) probabilities?
- 13.** What is Bayes' rule and what is it used for?
- 14.** Given  $P(A|B_i)$  and  $P(B_i)$ , how does one “reverse the probability” to determine  $P(B_i|A)$ ?
- 15.** What does it mean for two events  $A$  and  $B$  to be independent?
- 16.** What is  $P(A \cap B)$  when two events  $A$  and  $B$  are (i) mutually exclusive, and (ii) independent?

## EXERCISES

### Section 3.1

- 3.1** When two dice—one black with white dots, the other black with white dots—are tossed once, simultaneously, and the number of dots shown on each die's top face after coming to rest are recorded as an ordered pair  $(n_B, n_W)$ , where  $n_B$  is the number on the black die, and  $n_W$  the number on the white die,
- (i) identify the *experiment*, what constitutes a *trial*, the *outcomes*, and the *sample space*.
  - (ii) If the sum of the numbers on the two dice is  $S$ , i.e.,

$$S = n_B + n_W, \quad (3.69)$$

enumerate all the simple events associated with the observation  $S = 7$ .

- 3.2** In an opinion poll, 20 individuals selected at random from a group of college students are asked to indicate which of three options—*approve*, *disapprove*, *indifferent*—best matches their individual opinions of a new campus policy. Let  $n_0$  be the number of indifferent students,  $n_1$  the number that approve, and  $n_2$  the number that

disapprove, so that the outcome of one such opinion sample is the ordered triplet  $(n_0, n_1, n_2)$ . Write mathematical expressions in terms of the numbers  $n_0, n_1$ , and  $n_2$  for the following events:

- (i)  $A = \{\text{Unanimous support for the policy}\};$  and  $A^*$ , the complement of  $A$ .
- (ii)  $B = \{\text{More students disapprove than approve}\};$  and  $B^*$ .
- (iii)  $C = \{\text{More students are indifferent than approve}\};$
- (iv)  $D = \{\text{The majority of students are indifferent}\}.$

### Section 3.2

**3.3** Given the following two sets  $A$  and  $B$ :

$$A = \{x : x = 1, 3, 5, 7, \dots\} \quad (3.70)$$

$$B = \{x : x = 0, 2, 4, 6, \dots\} \quad (3.71)$$

find  $A \cup B$  and  $A \cap B$ .

**3.4** Let  $A_k = \{x : 1/(k+1) \leq x \leq 1\}$  for  $k = 1, 2, 3, \dots$ . Find the set  $B$  defined by:

$$B = A_1 \cup A_2 \cup A_3 \cup \dots = \bigcup_{i=1}^{\infty} A_i \quad (3.72)$$

**3.5** For sets  $A, B, C$ , subsets of the universal set  $\Omega$ , establish the following identities:

$$(A \cup B)^* = A^* \cap B^* \quad (3.73)$$

$$(A \cap B)^* = A^* \cup B^* \quad (3.74)$$

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \quad (3.75)$$

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C) \quad (3.76)$$

**3.6** For every pair of sets  $A, B$ , subsets of the *sample space*  $\Omega$  upon which the probability set function  $P(\cdot)$  has been defined, prove that:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (3.77)$$

**3.7** In a certain engineering research and development company, apart from the support staff which number 25, all other employees are either engineers or statisticians or both. The total number of employees (including the support staff) is 100. Of these, 50 are engineers, and 40 are statisticians; the number of employees that are both engineers and statisticians is not given. Find the probability that an employee chosen at random is *not* one of those classified as being both an engineer and a statistician.

### Section 3.3

**3.8** For every set  $A$ , let the set function  $Q(\cdot)$  be defined as follows:

$$Q(A) = \sum_A f(x) \quad (3.78)$$

where

$$f(x) = \left(\frac{2}{3}\right) \left(\frac{1}{3}\right)^x ; x = 0, 1, 2, \dots \quad (3.79)$$

If  $A_1 = \{x : x = 0, 1, 2, 3\}$  and  $A_2 = \{x : x = 0, 1, 2, 3, \dots\}$  find  $Q(A_1)$  and  $Q(A_2)$ .

**3.9** Let the sample space for a certain experiment be  $\Omega = \{\omega : 0 < \omega < \infty\}$ . Let  $A \subset \Omega$  represent the event  $A = \{\omega : 4 < \omega < \infty\}$ . If the probability set function  $P(A)$  is defined for any subset of the sample space according to:

$$P(A) = \int_A e^{-x} dx \quad (3.80)$$

evaluate  $P(A), P(A^*), P(A \cup A^*)$

**3.10** For the experiment of rolling two dice—one black with white dots, the other black with white dots—once, simultaneously, presented in Exercise 3.1, first obtain  $\Omega$ , the sample space, and, by assigning equal probability to each of the outcomes, determine the probability of the following events:

- (i)  $A = \{n_B + n_W = 7\}$ , i.e. the sum is 7;
- (ii)  $B = \{n_B < n_W\}$ ;
- (iii)  $B^*$ , the complement of  $B$ ;
- (iv)  $C = \{n_B = n_W\}$ , i.e. the two dice show the same number;
- (v)  $D = \{n_B + n_W = 5 \text{ or } 9\}$ .

**3.11** A black velvet bag contains three red balls and three green balls. Each experiment involves drawing two balls at once, simultaneously, and recording their colors,  $R$  for red, and  $G$  for green.

- (i) Obtain the sample space, assuming that balls of the same color are indistinguishable.
- (ii) Upon assigning equal probability to each element in the sample space, determine the probability of drawing two balls of different colors.
- (iii) If the balls are distinguishable and numbered from 1 to 6, and if the two balls are drawn sequentially, not simultaneously, now obtain the sample space and from this determine the probability of drawing two balls of different colors.

**3.12** An experiment is performed by selecting a card from an ordinary deck of 52 playing cards. The outcome,  $\omega$ , is the type of card chosen, classified as: “Ace,” “King,” “Queen,” “Jack,” and “others.” The random variable  $X(\omega)$  assigns the number 4 to the outcome if  $\omega$  is an “Ace;”  $X(\omega) = 3$  if the outcome is a “King”;  $X(\omega) = 2$  if the outcome is a “Queen,” and  $X(\omega) = 1$  if the outcome is a “Jack”;  $X(\omega) = 0$  for all other outcomes.

- (i) What is the space  $V$  of this random variable?
- (ii) If the probability set function  $P(\Gamma)$  defined on the subsets of the original sample space  $\Omega$  assigns a probability  $1/52$  to each of these outcomes, describe the induced probability set function  $P_X(A)$  induced on all the subsets of the space  $V$  by this random variable.
- (iii) Describe a physical (scientific or engineering) problem for which the above would be a good surrogate “model.”

**3.13** Obtain the sample space,  $\Omega$ , for the experiment involving tossing a fair coin 4 times. Upon assigning equal probability to each outcome, determine the probabilities of obtaining, 0, 1, 2, 3, or 4 heads. Confirm that your result is consistent with the

postulate that the probability model for this phenomenon is given by the probability distribution function:

$$f(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \quad (3.81)$$

where  $f(x)$  is the probability of obtaining  $x$  heads in  $n = 4$  tosses, and  $p = \frac{1}{2}$  is the probability of obtaining a head in a *single* toss of the coin. (See Chapter 8.)

**3.14** In the fall of 2007,  $k$  students born in 1989 attended an all-freshman introductory general engineering class at the University of Delaware. Confirm that if  $p$  is the probability that *at least two* of the students have the same birthday then:

$$1 - p = \frac{365!}{(365-k)!} \frac{1}{(365)^k} \quad (3.82)$$

Show that for a class with 23 or more students born in 1989, the probability of at least 2 students sharing the same birthday, is more than  $1/2$ , i.e., if  $k > 23$  then  $p > 1/2$ .

### Sections 3.4 and 3.5

**3.15** Six simple events, with probabilities  $P(E_1) = 0.11$ ;  $P(E_2) = P(E_5) = 0.20$ ;  $P(E_3) = 0.25$ ;  $P(E_4) = 0.09$ ;  $P(E_6) = 0.15$ , constitute the entire set of outcomes of an experiment. The following events are of interest:

$$A = \{E_1, E_2\}; B = \{E_2, E_3, E_4\}; C = \{E_5, E_6\}; D = \{E_1, E_2, E_5\}$$

Determine the following probabilities:

- (i)  $P(A), P(B), P(C), P(D)$ ;
- (ii)  $P(A \cup B), P(A \cap B); P(A \cup D), P(A \cap D); P(B \cup C), P(B \cap C)$ ;
- (iii)  $P(B|A), P(A|B); P(B|C), P(D|C)$

Which of the events  $A, B, C$  and  $D$  are mutually exclusive?

**3.16** Assuming that giving birth to a boy or a girl is equally likely, and further, that no multiple births have occurred, first, determine the probability of a family having three boys in a row. Now consider the conjecture (based on empirical data) that, for a family that has already had two boys in a row, the probability of having a third boy is 0.8. Under these conditions, what is now the probability of a family having three boys in a row?

**3.17** As a follow-up to the concept of independence of two events  $A$  and  $B$ ,

- Event  $A$  is said to be “attracted” to event  $B$  if

$$P(A|B) > P(A) \quad (3.83)$$

- Event  $A$  is said to be “repelled” by event  $B$  if

$$P(A|B) < P(A) \quad (3.84)$$

(Of course, when  $P(A|B) = P(A)$ , the two events have been previously identified as independent.) Establish the result that if  $B$  attracts  $A$ , then: (i)  $A$  attracts  $B$

(mutual attraction); and (ii)  $B^*$  repels  $A$ .

**3.18** Show that if  $A$  and  $B$  are independent, then  $A^*$  and  $B^*$  are also independent.

**3.19** Show that for two events  $A$  and  $B$ ,  $P(A \cup B | A \cap B) \leq P(A \cap B | A)$ . State the condition for equality.

**3.20** An exchange student from Switzerland, who is male, has been assigned to be your partner in an introductory psychology class. As part of a class assignment, he responds to your question about his family by stating only that he comes from a family of two children, without specifying whether he is the older or the younger. What is the probability that his sibling is female? Assume equal probability of having a boy or girl. Why does this result seem counterintuitive at first?

**3.21** A system consisting of two components  $A$  and  $B$  that are connected in *series* functions if *both* of them function. If  $P(A)$ , the probability that component  $A$  functions, is 0.99, and the probability that component  $B$  functions is 0.90, find the probability that this series system functions, assuming that whether one component functions or not is independent of the status of the other component. If these components are connected in *parallel*, the system fails (i.e., will *not* function) only if *both* components fail. Assuming independence, determine the probability that the parallel system functions. Which probability is higher and why is it reasonable to expect a higher probability from the system in question?

**3.22** The functioning status of a complex system that consists of several components arranged in series and parallel and with cross-links (i.e., whether the system functions or not) can be determined from the status of a “keystone” component,  $C_k$ . If the probability that the keystone component for a particular system functions is given as  $P(C_k) = 0.9$  and the probability that the system function when the keystone functions,  $P(S|C_k)$ , is given as 0.9, with the complementary probability that the system functions when the keystone *does not* function,  $P(S|C_k^*)$ , given as 0.8, find the unconditional probability,  $P(S)$ , that the system functions.

## APPLICATION PROBLEMS

**3.23** Patients suffering from manic depression and other similar disorders are sometimes treated with lithium, but the dosage must be monitored carefully because lithium toxicity, which is often fatal, can be difficult to diagnose. A new assay used to determine lithium concentration in blood samples is being promoted as a reliable way to diagnose lithium toxicity because the assay result is purported to correlate very strongly with toxicity.

A careful study of the relationship between this blood assay and lithium toxicity in 150 patients yielded results summarized in Table 3.3. Here  $A^+$  indicates high lithium concentrations in the blood assay and  $A^-$  indicates low lithium concentration;  $L^+$  indicates confirmed Lithium toxicity and  $L^-$  indicates *no* lithium toxicity.

(i) From these data, compute the following probabilities regarding the lithium toxicity status of a patient chosen at random:

**TABLE 3.3:** Lithium toxicity study results

| Assay | Lithium Toxicity |       | Total |
|-------|------------------|-------|-------|
|       | $L^+$            | $L^-$ |       |
| $A^+$ | 30               | 17    | 47    |
| $A^-$ | 21               | 82    | 103   |
| Total | 51               | 92    | 150   |

1.  $P(L^+)$ , the probability that the patient has lithium toxicity (regardless of the blood assay result);
  2.  $P(L^+|A^+)$ , the conditional probability that the patient has lithium toxicity given that the blood assay result indicates *high* lithium concentration. What does this value indicate about the potential benefit of having this assay result available?
  3.  $P(L^+|A^-)$  the conditional probability that the patient has lithium toxicity given that the blood assay result indicates *low* lithium concentration. What does this value indicate about the potential for missed diagnoses?
- (ii) Compute the following probabilities regarding the blood lithium assay:
1.  $P(A^+)$ , the (total) probability of observing high lithium blood concentration (regardless of actual lithium toxicity status);
  2.  $P(A^+|L^+)$  the conditional probability that the blood assay result indicates *high* lithium concentration given that the patient indeed has lithium toxicity. Why do you think that this quantity is referred to as the “sensitivity” of the assay, and what does the computed value indicate about the “sensitivity” of the particular assay in this study?
  3. From information about  $P(L^+)$  (as the “prior” probability of lithium toxicity) along with the just computed values of  $P(A^+)$  and  $P(A^+|L^+)$  as the relevant assay results, *now use Bayes’ Rule* to compute  $P(L^+|A^+)$  as the “posterior” probability of lithium toxicity *after* obtaining assay data, even though it has already been computed directly in (i) above.

**3.24** An experimental crystallizer produces five different polymorphs of the same crystal via mechanisms that are currently not well-understood. Types 1, 2 and 3 are approved for pharmaceutical application *A*; Types 2, 3 and 4 for a different application *B*; Type 5 is mostly unstable and has no known application. How much of each type is made in any batch varied randomly, but with the current operating procedure, 30% of the total product made by the crystallizer in a month is of Type 1; 20% is of Type 2, with the same percentage of Types 3 and 4; and 10% is of Type 5. Assuming that the polymorphs can be separated without loss,

- (i) Determine the probability of making product in a month that can be used for application *A*;
- (ii) Given a batch ready to be shipped for application *B*, what is the probabilities that any crystal selected at random is of Type 2? What is the probability that it is of Type 3 or Type 4. State any assumptions you may need to make.

- (iii) What is the probability that an order change to one for application  $A$  can be filled from a batch ready to be shipped for application  $B$ ?
- (iv) What is the converse probability that an order change to one for application  $B$  can be filled given a batch that is ready to be shipped for application  $A$ ?

**3.25** A test for a relatively rare disease involves taking from the patient an appropriate tissue sample which is then assessed for abnormality. A few sources of error are associated with this test. First, there is a small, but non-zero probability,  $\theta_s$ , that the tissue sampling procedure will miss abnormal cells primarily because these cells (at least in the earlier stages) being relatively few in number, are randomly distributed in the tissue and tend not to cluster. In addition, during the examination of the tissue sample itself, there is a probability,  $\theta_f$ , of failing to identify an abnormality when present; and a probability,  $\theta_m$ , of misclassifying a perfectly normal cell as abnormal.

If the proportion of the population with this disease who are subjected to this test is  $\theta_D$ ,

- (i) In terms of the given parameters, determine the probability that the test result is correct. (Hint: first compute the probability that the test result is incorrect, keeping in mind that the test may identify an abnormal cell incorrectly as normal, or a normal cell as abnormal.)
- (ii) Determine the probability of a false positive (i.e., returning an abnormality result when none exists).
- (iii) Determine the probability of a false negative (i.e., failing to identify an abnormality that is present).

**3.26** Repeat Problem 3.25 for the specific values of  $\theta_s = 0.1$ ;  $\theta_f = 0.05$ ;  $\theta_m = 0.1$  for a population in which 2% have the disease. A program sponsored by the Center for Disease Control (CDC) is to be aimed at reducing the number of false positives and/or false negatives by reducing one of the three probabilities  $\theta_s$ ,  $\theta_f$ , and  $\theta_m$ . Which of these parameters would you recommend and why?

**3.27** A manufacturer of flat-screen TVs purchases pre-cut glass sheets from three different manufacturers,  $M_1$ ,  $M_2$  and  $M_3$ , whose products are characterized in the TV manufacturer's "incoming material" quality control lab as "premier" grade,  $Q_1$ , "acceptable" grade,  $Q_2$ , and "marginal" grade,  $Q_3$ , on the basis of objective, measurable quality criteria, such as inclusions, warp, etc. Incoming glass sheets deemed unacceptable are rejected and returned to the manufacturer. An incoming batch of 425 *accepted* sheets has been classified by an automatic classifying system as shown in the table below.

| Manufacturer ↓ | Quality →        |                     |                   | Total |
|----------------|------------------|---------------------|-------------------|-------|
|                | Premier<br>$Q_1$ | Acceptable<br>$Q_2$ | Marginal<br>$Q_3$ |       |
| $M_1$          | 110              | 25                  | 15                | 150   |
| $M_2$          | 150              | 33                  | 2                 | 185   |
| $M_3$          | 76               | 13                  | 1                 | 90    |

If a sheet is selected at random from this batch,

- (i) Determine the probability that it is of "premier" grade; also determine the probability that it is *not* of "marginal" grade.

- (ii) Determine the probability that it is of “premier” grade given that it is from manufacturer  $M_1$ ; also determine the probability that it is of “premier” grade given that it is from either manufacturer  $M_2$  or  $M_3$ .
- (iii) Determine the probability that it is from manufacturer  $M_3$  given that it is of “marginal” grade; also determine the probability that it is from manufacturer  $M_2$  given that it is of “acceptable” grade.

**3.28** In a 1984 report<sup>2</sup>, the IRS published the information shown in the following table regarding 89.9 million federal tax returns it received, the income bracket of the filers, and the percentage audited.

| Income Bracket      | Number of filers (millions) | Percent Audited |
|---------------------|-----------------------------|-----------------|
| Below \$10,000      | 31.4                        | 0.34            |
| \$10,000 – \$24,999 | 30.7                        | 0.92            |
| \$25,000 – \$49,999 | 22.2                        | 2.05            |
| \$50,000 and above  | 5.5                         | 4.00            |

- (i) Determine the probability that a tax filer selected at random from this population would be audited.
- (ii) Determine the probability that a tax filer selected at random is in the \$25,000 – \$49,999 income bracket *and* was audited.
- (iii) If we know that a tax filer selected at random was audited, determine the probability that this person belongs in the “\$50,000 and above” income bracket.

---

<sup>2</sup>*Annual Report of Commissioner and Chief Counsel*, Internal Revenue Service, U.S. Department of Treasury, 1984, p 60.



# Chapter 4

## Random Variables and Distributions

|       |   |     |
|-------|---|-----|
| 4.1   | Introduction and Definition .....                 | 90  |
| 4.1.1 | Mathematical Concept of the Random Variable ..... | 90  |
| 4.1.2 | Practical Considerations .....                    | 93  |
| 4.1.3 | Types of Random Variables .....                   | 94  |
| 4.2   | Distributions .....                               | 95  |
| 4.2.1 | Discrete Random Variables .....                   | 95  |
| 4.2.2 | Continuous Random Variables .....                 | 98  |
| 4.2.3 | The Probability Distribution Function .....       | 100 |
| 4.3   | Mathematical Expectation .....                    | 102 |
| 4.3.1 | Motivating the Definition .....                   | 102 |
| 4.3.2 | Definition and Properties .....                   | 104 |
| 4.4   | Characterizing Distributions .....                | 107 |
| 4.4.1 | Moments of a Distributions .....                  | 107 |
| 4.4.2 | Moment Generating Function .....                  | 113 |
| 4.4.3 | Characteristic Function .....                     | 115 |
| 4.4.4 | Additional Distributional Characteristics .....   | 116 |
| 4.4.5 | Entropy .....                                     | 119 |
| 4.4.6 | Probability Bounds .....                          | 119 |
| 4.5   | Special Derived Probability Functions .....       | 122 |
| 4.5.1 | Survival Function .....                           | 122 |
| 4.5.2 | Hazard Function .....                             | 123 |
| 4.5.3 | Cumulative Hazard Function .....                  | 124 |
| 4.6   | Summary and Conclusions .....                     | 124 |
|       | REVIEW QUESTIONS .....                            | 126 |
|       | EXERCISES .....                                   | 129 |
|       | APPLICATION PROBLEMS .....                        | 133 |

*An idea, in the highest sense of that word,  
cannot be conveyed but by a symbol.*

S. T. Coleridge (1772-1834)

Even though the machinery of probability as presented thus far can already be used to solve some practical problems, its development is far from complete. In particular, with a sample space of raw outcomes that can be anything from attributes and numbers, to letters and other sundry objects, this most basic form of probability will be quite tedious and inefficient in dealing with general random phenomena. This chapter and the next one are devoted to completing the development of the machinery of probability with the introduction of the concept of the *random variable*, from which arises the probability distribution function—an efficient mathematical form for representing the ensemble behavior of general random phenomena. The emergence, properties and characteristics of the probability distribution function are discussed extensively in

this chapter for single dimensional random variables; the discussion is generalized to multi-dimensional random variables in the next chapter.

## 4.1 Introduction and Definition

### 4.1.1 Mathematical Concept of the Random Variable

In general, the sample space  $\Omega$  presented thus far may be quite tedious to describe and inefficient to analyze mathematically if its elements are not numbers. To facilitate mathematical analysis, it is desirable to find a means of converting this sample space into one with real numbers. This is achieved via the vehicle of the “random variable” defined as follows:

**Definition:** Given a random experiment with a sample space  $\Omega$ , let there be a function  $X$ , which assigns to each element  $\omega \in \Omega$ , one and only one real number  $X(\omega) = x$ . This function,  $X$ , is called a *random variable*.

Upon the introduction of this entity,  $X$ , the following happens (See Fig 4.1):

1.  $\Omega$  is mapped onto  $V$ , i.e.

$$V = \{x : X(\omega) = x, \omega \in \Omega\} \quad (4.1)$$

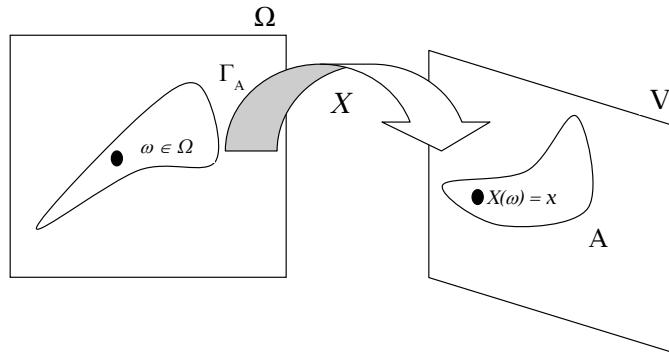
so that  $V$  is the set of all values  $x$  generated from  $X(\omega) = x$  for all elements  $\omega$  in the sample space  $\Omega$ ;

2. The probability set function encountered before,  $P$ , defined on  $\Omega$ , gives rise to another probability set function,  $P_X$ , defined on  $V$  and induced by  $X$ .  $P_X$  is therefore often referred to as an induced probability set function.

The role of  $P_X$  in  $V$  is identical to that of  $P$  in  $\Omega$ . Thus, for any arbitrary subsect  $A$  of  $V$ ,  $P_X(A)$  is the probability of event  $A$  occurring.

The primary question of practical importance may now be stated as follows: *How does one find  $P_X(A)$  in the new setting created by the introduction of the random variable  $X$ , given the original sample space  $\Omega$ , and the original probability set function  $P$  defined on it?*

The answer is to “go back” to what we know, i.e., to find that set  $\Gamma_A \subset \Omega$



**FIGURE 4.1:** The original sample space,  $\Omega$ , and the corresponding space  $V$  induced by the random variable  $X$

which corresponds to the set of values of  $\omega$  in  $\Omega$  that are mapped by  $X$  into  $A$ , i.e.

$$\Gamma_A = \{\omega : \omega \in \Omega \text{ and } X(\omega) \in A\} \quad (4.2)$$

Such a set  $\Gamma_A$  is called the “pre-image” of  $A$ , that set on the original sample space from which  $A$  is obtained when  $X$  is applied on its elements (see Fig 4.1). We now simply define

$$P_X(A) = P(\Gamma_A) \quad (4.3)$$

since, by definition of  $\Gamma_A$ ,

$$P\{X(\omega) \in A\} = P\{\omega \in \Gamma_A\} \quad (4.4)$$

from where we see how  $X$  induces  $P_X(\cdot)$  from the known  $P(\cdot)$ . It is easy to show that the induced  $P_X$  is an authentic probability set function in the spirit of Kolmogorov’s axioms.

#### Remarks:

1. The random variable is  $X$ ; the value it takes is the real number  $x$ . The one is a completely different entity from the other.
2. The expression  $P(X = x)$  will be used to indicate “the probability that the application of the random variable  $X$  results in an outcome with assigned value  $x$ ;” or, more simply, “the probability that the random variable  $X$  takes on a *particular* value  $x$ ”. As such, “ $X = x$ ” should not be confused with the familiar arithmetic statement of equality or equivalence.
3. In many instances, the starting point is the space  $V$  and not the tedious sample space  $\Omega$ , with  $P_X(\cdot)$  already defined so that there is no further need for reference to a  $P(\cdot)$  defined on  $\Omega$ .

Let us illustrate these concepts with some examples.

**Example 4.1 RANDOM VARIABLE AND INDUCED PROBABILITY FUNCTION FOR COIN TOSS EXPERIMENT**

The experiment in Example 3.1 in Chapter 3 involved tossing a coin 3 times and recording the number of observed heads and tails. From the sample space  $\Omega$  obtained there, define a random variable  $X$  as *the total number of tails obtained in the 3 tosses*. (1) Obtain the new space  $V$  and, (2) if  $A$  is the event that  $X = 2$ , determine the probability of this event's occurrence.

**Solution:**

(1) Recall from Example 3.1 that the sample space  $\Omega$  is given by

$$\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\} \quad (4.5)$$

consisting of all possible 8 outcomes, represented respectively, as  $\omega_i; i = 1, 2, \dots, 8$ , i.e.

$$\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6, \omega_7, \omega_8\}. \quad (4.6)$$

This is clearly one of the “tedious” types, not as conveniently amenable to mathematical manipulation. And now, by definition of  $X$ , we see that  $X(\omega_1) = 0; X(\omega_2) = X(\omega_3) = X(\omega_4) = 1; X(\omega_5) = X(\omega_6) = X(\omega_7) = 2; X(\omega_8) = 3$ , from where we now obtain the space  $V$  as:

$$V = \{0, 1, 2, 3\} \quad (4.7)$$

since these are all the possible values that  $X$  can take.

(2) To obtain  $P_X(A)$ , first we find  $\Gamma_A$ , the pre-image of  $A$  in  $\Omega$ . In this case,

$$\Gamma_A = \{\omega_5, \omega_6, \omega_7\} \quad (4.8)$$

so that upon recalling the probability set function  $P(\cdot)$  generated in Chapter 3 on the assumption of equiprobable outcomes, we obtain  $P(\Gamma_A) = 3/8$ , hence,

$$P_X(A) = P(\Gamma_A) = 3/8 \quad (4.9)$$

The next two examples illustrate sample spaces that occur naturally in the form of  $V$ .

**Example 4.2 SAMPLE SPACE FOR SINGLE DIE TOSS EXPERIMENT**

Consider an experiment in which a single die is thrown and the outcome is the *number* that shows up on the die's top face when it comes to rest. Obtain the sample space of all possible outcomes.

**Solution:**

The required sample space is the set  $\{1, 2, 3, 4, 5, 6\}$ , since this set of numbers is an exhaustive collection of all the possible outcomes of this experiment. Observe that this is a set of real numbers, so that it is already in the form of  $V$ . We can therefore define a probability set function directly on it, with no further need to obtain a separate  $V$  and an induced  $P_X(\cdot)$ .

Strictly speaking, this last example did involve an “implicit” application of a random variable, if we acknowledge that the “primitive” outcomes for the die toss experiment are actually dots on the top face of the die. However, by pre-specifying the outcome as a “count” of the dots shown on the resting top face, we simply skipped a step and went straight to the result of the application of the random variable transforming the dots to the count. The next example also involves similar die tosses, but in this case, the application of the random variable is explicit, following the “implicit” one that automatically produces numbers as the de-facto outcomes.

**Example 4.3 SAMPLE SPACE AND RANDOM VARIABLE FOR DOUBLE DICE TOSS EXPERIMENT**

Consider an experiment in which *two* dice are thrown at the same time, and the outcome is an ordered pair of the *numbers* that show up on each die’s top face after coming to rest. Assume that we are careful to specify and distinguish a “first” die (black die with white spots) from the “second” (white die with black spots) (1) Obtain the sample space of all possible outcomes. (2) Define a random variable  $X$  as the *sum* of numbers that show up on the two dice; obtain the new space  $V$  arising as a result of the application of this random variable.

**Solution:**

(1) The original sample space,  $\Omega$ , of the raw outcomes, is given by

$$\Omega = \{(1, 1), (1, 2), \dots, (1, 6); (2, 1), (2, 2), \dots, (6, 6)\} \quad (4.10)$$

a set of the 36 ordered pairs of all possible outcomes  $(n_1, n_2)$ , where  $n_1$  is the number showing up on the face of the first die, and  $n_2$  is the number on the second die. (Had it not been possible to distinguish a “first die” from the “second”, outcomes such as  $(2, 1)$  and  $(1, 2)$  could not have been distinguishable, and  $\Omega$  will contain only 21 elements, the 6 diagonal and one set of the 15 off-diagonal elements of the  $6 \times 6$  matrix of ordered pairs.)

The elements of this set are clearly real numbers — the 2-dimensional kind — and are already amenable to mathematical manipulation. The definition of a random variable  $X$  in this case is therefore not for purposes of converting  $\Omega$  to a more mathematically convenient form; the random variable definition is a reflection of what aspect of the experiment is of interest to us.

(2) By the definition of  $X$ , we see that the required space  $V$  is:

$$V = \{2, 3, 4, \dots, 12\} \quad (4.11)$$

a set containing 11 elements, a collection of all the possible values that  $X$  can take in this case.

As an exercise, (see Exercise 4.7) the reader should compute the probability  $P_X(A)$  of the event  $A$  that  $X = 7$ , assuming equiprobable outcomes for each die toss.

### 4.1.2 Practical Considerations

Rigor and precision are intrinsic to mathematics and mathematical analysis; without the former, the latter simply cannot exist. Such is the case with the mathematical concept of the random variable as we have just presented it: rigor demands that  $X$  be specified in this manner, as a function through whose agency *each* element of the sample space of an experiment becomes associated with an unambiguous numerical value. As illustrated in Fig 4.1,  $X$  therefore appears as a mapping from one space,  $\Omega$ , that can contain all sorts of raw objects, into one that is more conducive to mathematical analysis,  $V$ , containing only real numbers. Such a formal definition of the random variable tends to appear stiff, and almost sterile; and those encountering it for the first time may be unsure of what it really means in practice.

As a practical matter, the random variable may be considered (informally) as an experimental outcome whose numerical value is subject to random variations with each exact replicate performance (trial) of the experiment. Thus, for example, with the “three coin-toss” experiment discussed earlier, by specifying the outcome of interest as the total number of tails observed, we see right away that the implied random variable can take on numerical values 0, 1, 2, or 3, even though the raw outcomes will consist of *T*s and *H*s; also what value the random variable takes is subject to random variation each time the experiment is performed. In the same manner, we see that in attempting to determine the temperature of an equilibrium mixture of ice and water, the observed temperature measurement in  $^{\circ}\text{C}$  takes on numerical values that vary randomly around the number 0.

### 4.1.3 Types of Random Variables

A random variable can be either *discrete* or *continuous*, as determined by the nature of the space  $V$ . For a discrete random variable, the space  $V$  consists of isolated points—“isolated” in the sense that, on the real line, every neighborhood of each point contains no other point of  $V$ . For instance, in Example 4.1 above, the random variable  $X$  can only take values 0, 1, 2, or 3; it is therefore a discrete random variable.

On the other hand, the space  $V$  associated with a continuous random variable consists of an interval of the real line, or, in higher dimensions, a set of intervals. For example, let  $\Omega$  be defined as:

$$\Omega = \{\omega : -1 \leq \omega \leq 1\} \quad (4.12)$$

If we define a random variable  $X$  as:

$$X(\omega) = 1 - |\omega|, \quad (4.13)$$

observe that the random variable space  $V$  in this case is given by:

$$V = \{x : 0 \leq x \leq 1\}. \quad (4.14)$$

This is an example of a continuous random variable.

Random variables can also be defined in higher dimensions. For example, given a sample space  $\Omega$  with a probability set function  $P(\cdot)$  defined on its subsets, a two-dimensional random variable is a function defined on  $\Omega$  which assigns one and only one ordered number pair  $(X_1(\omega), X_2(\omega))$  to each element  $\omega \in \Omega$ . Associated with this random variable is a space  $V$  and a probability set function  $P_X$  induced by  $X = (X_1, X_2)$ , where  $V$  is defined as:

$$V = \{(x_1, x_2) : X_1(\omega) = x_1, X_2(\omega) = x_2; \omega \in \Omega\} \quad (4.15)$$

The following is a simple example of a two-dimensional random variable.

**Example 4.4 A 2-DIMENSIONAL RANDOM VARIABLE AND ITS SAMPLE SPACE**

Revisit Example 4.1 and the problem discussed therein involving tossing a coin 3 times and recording the number of observed heads and tails; define the following 2-dimensional random variable:  $X_1$  = total number of tails;  $X_2$  = total number of heads. Obtain the sample space  $V$  associated with this random variable.

**Solution:**

The required sample space in this case is:

$$V = \{(0, 3), (1, 2), (2, 1), (3, 0)\}. \quad (4.16)$$

Note that the two component random variables  $X_1$  and  $X_2$  are not independent since their sum,  $X_1 + X_2$ , by virtue of the experiment, is constrained to equal 3 always.

What is noted briefly here for two dimensions can be generalized to  $n$ -dimensions, and the next chapter is devoted entirely to a discussion of multi-dimensional random variables.

## 4.2 Distributions

### 4.2.1 Discrete Random Variables

Let us return once more to Example 4.1 and, this time, for each element of  $V$ , compute  $P(X = x)$ , and denote this by  $f(x)$ ; i.e.

$$f(x) = P(X = x) \quad (4.17)$$

Observe that  $P(X = 0) = P(\Gamma_0)$  where  $\Gamma_0 = \{\omega_1\}$ , so that

$$f(0) = P(X = 0) = 1/8 \quad (4.18)$$

Similarly, we obtain  $P(X = 1) = P(\Gamma_1)$  where  $\Gamma_1 = \{\omega_2, \omega_3, \omega_4\}$ , so that:

$$f(1) = P(X = 1) = 3/8 \quad (4.19)$$

Likewise,

$$f(2) = P(X = 2) = 3/8 \quad (4.20)$$

$$f(3) = P(X = 3) = 1/8 \quad (4.21)$$

This function,  $f(x)$ , indicates how the probabilities are distributed over the entire random variable space.

Of importance also is a different, but related, function,  $F(x)$ , defined as:

$$F(x) = P(X \leq x) \quad (4.22)$$

the probability that the random variable  $X$  takes on values less than or equal to  $x$ . For the specific example under consideration, we have:  $F(0) = P(X \leq 0) = 1/8$ . As for  $F(1) = P(X \leq 1)$ , since the event  $A = \{X \leq 1\}$  consists of two mutually exclusive elementary events  $A_0 = \{X = 0\}$  and  $A_1 = \{X = 1\}$ , it then follows that:

$$F(1) = P(X \leq 1) = P(X = 0) + P(X = 1) = 1/8 + 3/8 = 4/8 \quad (4.23)$$

By similar arguments, we obtain:

$$F(2) = P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = 7/8 \quad (4.24)$$

$$\begin{aligned} F(3) = P(X \leq 3) &= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) \\ &= 8/8 \end{aligned} \quad (4.25)$$

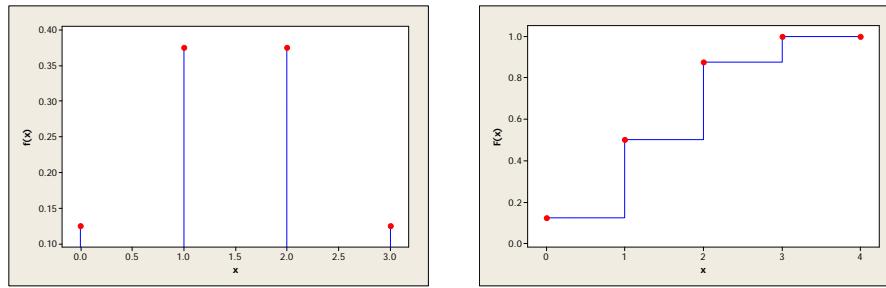
These results are tabulated in Table 4.1.

**TABLE 4.1:**  
 $f(x)$  and  $F(x)$  for  
 the “three coin-toss”  
 experiments of  
 Example 4.1

| $x$ | $f(x)$ | $F(x)$ |
|-----|--------|--------|
| 0   | 1/8    | 1/8    |
| 1   | 3/8    | 4/8    |
| 2   | 3/8    | 7/8    |
| 3   | 1/8    | 8/8    |

The function,  $f(x)$ , is referred to as the *probability distribution function* (pdf), or sometimes as the probability mass function;  $F(x)$  is known as the *cumulative distribution function*, or sometimes simply as the distribution function.

Note, once again, that  $X$  can assume only a finite number of discrete values, in this case, 0, 1, 2, or 3; it is therefore a discrete random variable, and both  $f(x)$  and  $F(x)$  are discrete functions. As shown in Fig 4.2,  $f(x)$  is characterized by non-zero “spikes” at values of  $x = 0, 1, 2$  and  $3$ , and  $F(x)$  by the indicated “staircase” form.



**FIGURE 4.2:** Probability distribution function,  $f(x)$ , and cumulative distribution function,  $F(x)$ , for 3-coin toss experiment of Example 4.1

Let  $x_0 = 0, x_1 = 1, x_2 = 2, x_3 = 3$ ; then

$$P(X = x_i) = f(x_i) \text{ for } i = 0, 1, 2, 3 \quad (4.26)$$

with  $f(x_i)$  given explicitly as:

$$f(x_i) = \begin{cases} 1/8; & x_0 = 0 \\ 3/8; & x_1 = 1 \\ 3/8; & x_2 = 2 \\ 1/8; & x_3 = 3 \end{cases} \quad (4.27)$$

and the two functions in Table 4.1 are related explicitly according to the following expression:

$$F(x_i) = \sum_{j=0}^i f(x_j) \quad (4.28)$$

We may now also note the following about the function  $f(x_i)$ :

- $f(x_i) > 0; \forall x_i$

- $\sum_{i=0}^3 f(x_i) = 1$

These ideas may now be generalized beyond the specific example used above.

**Definition:** Let there exist a sample space  $\Omega$  (along with a probability set function,  $P$ , defined on its subsets), and a random variable  $X$ , with an attendant random variable space  $V$ : a function  $f$  defined on  $V$  such that:

1.  $f(x) \geq 0; \forall x \in V;$
2.  $\sum_x f(x) = 1; \forall x \in V;$
3.  $P_X(A) = \sum_{x \in A} f(x);$  for  $A \subset V$  (and when  $A$  contains the single element  $x_i$ ,  $P_X(X = x_i) = f(x_i)$ )

is called a probability distribution function of the random variable  $X$ .

Upon comparing these formal statements regarding  $f(x)$  to the 3 axioms of Kolmogorov (regarding the probability set function  $P$  defined on  $\Omega$ ) given earlier in Chapter 3, we readily see that these are the same concepts extended from  $\Omega$  to  $V$  for the random variable  $X$ .

#### 4.2.2 Continuous Random Variables

For the continuous random variable  $X$ , because it takes on a continuum of values, not discrete points as with the discrete counterpart, the concepts presented above are modified as follows, primarily by replacing sums with integrals:

**Definition:** The function  $f$  defined on the space  $V$  (whose elements consist of segments of the real line) such that:

1.  $f(x) \geq 0; \forall x \in V;$
2.  $f$  has at most a finite number of discontinuities in every finite interval;
3. The (Riemann) integral,  $\int_V f(x)dx = 1;$
4.  $P_X(A) = \int_A f(x)dx;$  for  $A \subset V$

is called a probability *density* function of the continuous random variable  $X$ .

(The second point above, unnecessary for the discrete case, is a mathematical “fine point” needed to safeguard against pathological situations where the

probability “measure” becomes undefined; it is hardly ever an issue in most practical applications.)

In this case, the expression for the cumulative distribution function,  $F(x)$ , corresponding to that in Eq (4.28), is:

$$F(x_i) = P(X \leq x_i) = \int_{-\infty}^{x_i} f(x)dx \quad (4.29)$$

from where we may now observe that when  $F(x)$  possesses a derivative,

$$\frac{dF(x)}{dx} = f(x) \quad (4.30)$$

This  $f(x)$  is the continuous counterpart of the discrete  $f(x)$  encountered earlier; but rather than express the probability that  $X$  takes on a particular point value  $x_i$  (as in the discrete case), the continuous  $f(x)$  expresses a measure of the probability that  $X$  lies in the infinitesimal interval between  $x_i$  and  $x_i + dx$ . Observe, from item 4 in the definition given above, that:

$$P(x_i \leq X \leq x_i + dx) = \int_{x_i}^{x_i + dx} f(x)dx \approx f(x_i)dx \quad (4.31)$$

for a very small interval size  $dx$ .

In general, because the event  $A = \{X \leq x + dx\}$  can be decomposed into 2 mutually exclusive events  $B = \{x \leq X\}$  and  $C = \{x \leq X \leq x + dx\}$ , so that:

$$\{X \leq x + dx\} = \{x \leq X\} \cup \{x \leq X \leq x + dx\}, \quad (4.32)$$

we see that:

$$\begin{aligned} P(X \leq x + dx) &= P(x \leq X) + P(x \leq X \leq x + dx) \\ F(x + dx) &= F(x) + P(x \leq X \leq x + dx) \end{aligned} \quad (4.33)$$

and therefore:

$$P(x \leq X \leq x + dx) = F(x + dx) - F(x) \quad (4.34)$$

which, upon introducing Eq (4.31) for the LHS, dividing by  $dx$ , and taking limits as  $dx \rightarrow 0$ , yields:

$$\lim_{dx \rightarrow 0} \left[ \frac{F(x + dx) - F(x)}{dx} \right] = \frac{dF(x)}{dx} = f(x) \quad (4.35)$$

establishing Eq (4.30).

In general, we can use Eq (4.29) to establish that, for any arbitrary  $b \geq a$ ,

$$P(a \leq X \leq b) = \int_a^b f(x)dx = F(b) - F(a). \quad (4.36)$$

For the sake of completeness, we note that  $F(x)$ , the cumulative distribution function, is actually the more fundamental function for determining probabilities. This is because, regardless of whether  $X$  is continuous or discrete,  $F(\cdot)$  can be used to determine all desired probabilities. Observe from the foregoing discussion that the expression

$$P(a_1 < X \leq a_2) = F(a_2) - F(a_1) \quad (4.37)$$

will hold true whether  $X$  is continuous or discrete.

From now on in this book, we will simply talk of the “probability distribution function” (or “pdf” for short) for all random variables  $X$  (continuous or discrete) and mean by this, *probability distribution function* if  $X$  is discrete, and *probability density function* if  $X$  is continuous and expect that the context will make clear what we mean.

### 4.2.3 The Probability Distribution Function

We have now seen that the pdf  $f(x)$  (or equivalently, the cdf  $F(x)$ ) is the function that indicates how the probabilities of occurrence of various outcomes and events arising from the random phenomenon in question are distributed over the entire space of the associated random variable  $X$ .

Let us return once more to the “three coin-toss” example: we understand that the random phenomenon in question is such that we cannot predict, *à priori*, the specific outcome of each experiment; but from the ensemble aggregate of all possible outcomes, we have been able to characterize, with  $f(x)$ , the “behavior” of an associated random variable of interest,  $X$ , the total number of tails obtained in the experiment. (Note that other random variables could also be defined for this experiment: for example, the total number of heads, or the number of tosses until the appearance of the first head, etc.) What Table 4.1 provides is a complete description of the probability of occurrence for the entire collection of all possible events associated with this random variable—a description that can now be used to analyze the particular random phenomenon of the total number of tails observed when a coin is tossed three times.

For instance, the pdf  $f(x)$  indicates that, even though we cannot predict a specific outcome precisely, we now know that after each experiment, observing “no tails” ( $X = 0$ ) is just as likely as observing “all tails” ( $X = 3$ ), each with a probability of  $1/8$ . Also, observing “two tails” is just as likely as observing “one tail”, each with a probability of  $3/8$ , so that these latter group of events are three times as likely as the former group of events. Note the symmetry of the distribution of probabilities indicated by  $f(x)$  for this particular random phenomenon.

It turns out that these specific results can be generalized for the class of random phenomena to which the “three coin-toss” example belongs — a class characterized by the following features:

1. each experiment involves  $n$  identical trials (e.g. coin tosses, or number of fertilized embryos transferred in an in-vitro fertilization (IVF) treatment cycle, etc), and *each* trial can produce only two mutually exclusive outcomes: “S” (success) or “F” (failure);
2. the probability of “success” in each trial is  $p$ ; and,
3. the outcome of interest,  $X$ , is the total of “successes” observed (e.g. tails in coin tosses, live births in IVF, etc).

As we show a bit later, the pdf characterizing this family of random phenomena is given by:

$$f(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}; x = 0, 1, 2, \dots, n \quad (4.38)$$

The results in Table 4.1 are obtained for the special case  $n = 3$ ;  $p = 0.5$ .

Such functions as these provide convenient and compact mathematical representations of the desired ensemble behavior of random variables; they constitute the centerpiece of the probabilistic framework — the fundamental tool used for analyzing random phenomena.

We have, in fact, already encountered in earlier chapters, several actual pdfs for some real-world random variables. For example, we had stated in Chapter 1 (thus far without justification) that the continuous random variable representing the yield obtained from the example manufacturing processes has the pdf:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}; -\infty < x < \infty \quad (4.39)$$

We are able to use this pdf to compute the probabilities of obtaining yields in various intervals on the real line for the two contemplated processes, once the parameters  $\mu$  and  $\sigma$  are specified for each process.

We had also stated in Chapter 1 that, for the (discrete) random variable  $X$  representing the number of “inclusions” found on the manufactured glass sheet, the pdf is:

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}; x = 0, 1, 2, \dots \quad (4.40)$$

from which, again, given a specific value for the parameter  $\lambda$ , we are able to compute the probabilities of finding any given number of “inclusions” on any selected glass sheet. And in Chapter 2, we showed, using chemical engineering principles, that the pdf for the (continuous) random variable  $X$ , representing the residence time in an ideal CSTR, is given by:

$$f(x) = \frac{1}{\tau} e^{-x/\tau}; 0 < x < \infty \quad (4.41)$$

an expression that is used in practice for certain aspects of chemical reactor design.

These pdfs are all ideal models of the random variability associated with each of the random variables in question; they make possible rigorous and precise mathematical analyses of the ensemble behavior of the respective random phenomena. Such mathematical representations are systematically derived for actual, specific real-world phenomena of practical importance in Part III, where the resulting pdfs are also discussed and analyzed extensively.

The rest of this chapter is devoted to taking a deeper look at the fundamental characteristics and general properties of the pdf,  $f(x)$ , for single-dimensional random variables; the next chapter is devoted to a parallel treatment for multi-dimensional random variables.

### 4.3 Mathematical Expectation

We begin our investigations into the fundamental characteristics of a random variable,  $X$ , and its pdf,  $f(x)$ , with one of the most important: the “mathematical expectation” or “expected value.” As will soon become clear, the concept of expectations of random variables (or functions of random variables) is of significant practical importance; but before giving a formal definition, we first provide a motivation and an illustration of the concept.

#### 4.3.1 Motivating the Definition

Consider a game where each turn involves a player drawing a ball at random from a black velvet bag containing 9 balls, identical in every way except that 5 are red, 3 are blue and one is green. The player receives \$1.00 for drawing a red ball, \$4.00 for a blue ball, and \$10.00 for the green ball, but each turn at the game costs \$4.00 to play. The question is: *Is this game worth playing?*

The primary issue, of course, is the random variation in the color of the drawn ball each time the game is played. Even though simple and somewhat artificial, this example provides a perfect illustration of how to solve problems involving random phenomena using the probabilistic framework.

To arrive at a rational decision regarding whether to play this game or not, we proceed as follows, noting first the following characteristics of the phenomenon in question:

- *Experiment:* Draw a ball at random from a bag containing 9 balls composed as given above; note the color of the drawn ball, then replace the ball;
- *Outcome:* The color of the drawn ball:  $R$  = Red;  $B$  = Blue;  $G$  = Green.

#### Probabilistic Model Development

**TABLE 4.2:**

The pdf  $f(x)$  for  
the ball-drawing  
game

| $x$ | $f(x)$ |
|-----|--------|
| 1   | 5/9    |
| 4   | 3/9    |
| 10  | 1/9    |

From the problem definition, we see that the sample space  $\Omega$  is given by:

$$\Omega = \{R, R, R, R, R, B, B, B, G\} \quad (4.42)$$

The random variable,  $X$ , is clearly the monetary value assigned to the outcome of each draw; i.e. in terms of the formal definition,  $X$  assigns the real number 1 to  $R$ , 4 to  $B$ , and 10 to  $G$ . (Informally, we could just as easily say that  $X$  is the amount of money received upon each draw.) The random variable space  $V$  is therefore given by:

$$V = \{1, 4, 10\} \quad (4.43)$$

And now, since there is no reason to think otherwise, we assume that each outcome is equally probable, in which case the probability distribution for the random variable  $X$  is obtained as follows:

$$P_X(X = 1) = P(R) = 5/9 \quad (4.44)$$

$$P_X(X = 4) = P(B) = 3/9 \quad (4.45)$$

$$P_X(X = 10) = P(G) = 1/9 \quad (4.46)$$

so that  $f(x)$ , the pdf for this discrete random variable, is as shown in the Table 4.2, or, mathematically as:

$$f(x_i) = \begin{cases} 5/9; & x_1 = 1 \\ 3/9; & x_2 = 4 \\ 1/9; & x_3 = 10 \\ 0; & \text{otherwise} \end{cases} \quad (4.47)$$

This is an ideal model of the random phenomenon underlying this game; it will now be used to analyze the problem and to decide rationally whether to play the game or not.

### Using the Model

We begin by observing that this is a case where it is possible to repeat the experiment a large number of times; in fact, this is precisely what the person setting up the game wants each player to do: play the game repeatedly! Thus, if the game is played a very large number of times, say  $n$ , it is reasonable from the model to expect  $5n/9$  red ball draws,  $3n/9$  blue ball draws, and  $n/9$  green

ball draws; the corresponding financial returns will be  $\$(5n/9)$ ,  $\$(4 \times 3n/9)$ , and  $\$(10 \times n/9)$ , respectively, in each case.

Observe now that after  $n$  turns at the game, we would expect the total financial returns in dollars, say  $R_n$ , to be:

$$R_n = \left( 1 \times \frac{5n}{9} + 4 \times \frac{3n}{9} + 10 \times \frac{n}{9} \right) = 3n \quad (4.48)$$

These results are summarized in Table 4.3.

**TABLE 4.3:** Summary analysis for the ball-drawing game

| Ball Color | Expected # of times drawn (after $n$ trials) | Financial returns per draw | Expected financial returns (after $n$ trials) |
|------------|--|----------------------------|---|
| Red        | $5n/9$                                       | 1                          | $\$5n/9$                                      |
| Blue       | $3n/9$                                       | 4                          | $\$12n/9$                                     |
| Green      | $n/9$  | 10                         | $\$10n/9$                                     |
| Total      |  |                            | 3n  |

In the meantime, the total cost  $C_n$ , the amount of money, in dollars, paid out to play the game, would have been  $4n$ . On the basis of these calculations, therefore, the expected net gain (in dollars) after  $n$  trials,  $G_n$ , is given by

$$G_n = R_n - C_n = -n \quad (4.49)$$

indicating a net *loss* of  $\$n$ , so that the rational decision is *not* to play the game. (The house always wins!)

Eq (4.48) implies that the expected return *per draw* will be:

$$\frac{R_n}{n} = \left( 1 \times \frac{5}{9} + 4 \times \frac{3}{9} + 10 \times \frac{1}{9} \right) = 3, \quad (4.50)$$

a sum of all possible values of the random variable  $X$ , weighted by their corresponding probabilities, i.e. from Eq (4.47),

$$\frac{R_n}{n} = \sum_{i=1}^3 x_i f(x_i) \quad (4.51)$$

This quantity is known as the *expected value*, or the *mathematical expectation* of the random variable  $X$  — a weighted average of the values taken by  $X$  with the respective probabilities of obtaining each value as the weights.

We are now in a position to provide a formal definition of the mathematical expectation.

### 4.3.2 Definition and Properties

The expected value, or mathematical expectation, of a random variable, denoted by  $E(X)$ , is defined for a discrete random variable as:

$$E(X) = \sum_i x_i f(x_i) \quad (4.52)$$

and for a continuous random variable,

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx \quad (4.53)$$

provided that the following conditions hold:

$$\sum_i |x_i| f(x_i) < \infty \quad (4.54)$$

(known as “absolute convergence”) for discrete  $X$ , and

$$\int_{-\infty}^{\infty} |x| f(x) dx < \infty \quad (4.55)$$

(absolute integrability) for continuous  $X$ . If these conditions are *not* satisfied, then  $E(X)$  *does not exist* for  $X$ .

If the pdf is interpreted as an assignment of weights to point values of a discrete random variable, or intervals of a continuous random variable, then observe that  $E(X)$  is that value of the random variable that is the “center of gravity” of the distribution.

Some important points to note about the mathematical expectation:

1.  $E(X)$  is *not* a random variable; it is an exactly defined real number;
2. When  $X$  has units,  $E(X)$  has the same units as  $X$ ;
3.  $E(X)$  is often called the mean value of the random variable (or equivalently, of its distribution  $f(x)$ ), represented as  $\mu(X)$  or simply  $\mu$ , thus:

$$E(X) = \mu(X) \quad (4.56)$$

#### Example 4.5 EXPECTED VALUE OF TWO DISCRETE RANDOM VARIABLES

(1) Find the expected value of the random variable,  $X$ , the total number of tails observed in the three coin-toss experiment, whose pdf  $f(x)$  is given in Table 4.1 and in Eq (4.27).

(2) Find the expected value of the random variable,  $X$ , the financial returns on the ball-draw game, whose pdf  $f(x)$  is given in Eq (4.47).

**Solution:**

(1) From the definition of  $E(X)$ , we have in this case,

$$E(X) = (0 \times 1/8 + 1 \times 3/8 + 2 \times 3/8 + 3 \times 1/8) = 1.5 \quad (4.57)$$

indicating that with this experiment, the expected, or *average*, number of tails per toss is 1.5, which makes perfect sense.

(2) The expected financial return for the ball-draw game is obtained formally from Eq (4.47) as:

$$E(X) = (1 \times 5/9 + 4 \times 3/9 + 10 \times 1/9) = 3.0 \quad (4.58)$$

as we had obtained earlier.

**Example 4.6 EXPECTED VALUE OF TWO CONTINUOUS RANDOM VARIABLES**

(1) Find the expected value of the random variable,  $X$ , whose pdf  $f(x)$  is given by:

$$f(x) = \begin{cases} \frac{1}{2}x; & 0 < x < 2 \\ 0; & \text{otherwise} \end{cases} \quad (4.59)$$

(2) Find the expected value of the random variable,  $X$ , the residence time in a CSTR, whose pdf  $f(x)$  is given in Eq (4.41).

**Solution:**

(1) First, we observe that Eq (4.59) is a legitimate pdf because

$$\int_{-\infty}^{\infty} f(x)dx = \int_0^2 \frac{1}{2}x dx = \frac{1}{4}x^2 \Big|_0^2 = 1 \quad (4.60)$$

and, by definition,

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \frac{1}{2} \int_0^2 x^2 dx = \frac{1}{6}x^3 \Big|_0^2 = \frac{4}{3} \quad (4.61)$$

(2) In the case of the residence time,

$$E(X) = \int_{-\infty}^{\infty} \frac{1}{\tau} xe^{-x/\tau} dx = \frac{1}{\tau} \int_0^{\infty} xe^{-x/\tau} dx \quad (4.62)$$

since the random variable  $X$ , residence time, takes no negative values. Upon integrating the RHS by parts, we obtain,

$$E(X) = -xe^{-x/\tau} \Big|_0^{\infty} + \int_0^{\infty} e^{-x/\tau} dx = 0 - \tau e^{-x/\tau} \Big|_0^{\infty} = \tau \quad (4.63)$$

indicating that the expected, or average, residence time is the reactor parameter  $\tau$ , providing justification for why this parameter is known in chemical reactor design as the “mean residence time.”

An important property of the mathematical expectation of a random variable  $X$  is that for any function of this random variable, say  $G(X)$ ,

$$E[G(X)] = \begin{cases} \sum_i G(x_i)f(x_i); & \text{for discrete } X \\ \int_{-\infty}^{\infty} G(x)f(x)dx; & \text{for continuous } X \end{cases} \quad (4.64)$$

provided that the conditions of absolute convergence and absolute integrability stated earlier for  $X$  in Eqs (4.54) and (4.55), respectively, hold for  $G(X)$ .

In particular, if  $G(X)$  is a linear function, say for example,

$$G(X) = c_1X + c_2 \quad (4.65)$$

where  $c_1$  and  $c_2$  are constants, then from Eq (4.64) above, in the discrete case, we have that:

$$\begin{aligned} E(c_1X + c_2) &= \sum_i (c_1x_i + c_2)f(x_i) \\ &= c_1 \sum_i x_i f(x_i) + c_2 \sum_i f(x_i) \\ &= c_1E(X) + c_2 \end{aligned} \quad (4.66)$$

so that:

$$E(c_1X + c_2) = c_1E(X) + c_2 \quad (4.67)$$

Thus, treated like an operator,  $E(\cdot)$  is a linear operator. Similar arguments follow for the continuous case, replacing sums with appropriate integrals (see end-of-chapter Exercise 4.12).

## 4.4 Characterizing Distributions

One of the primary utilities of the result in Eq (4.64) is for obtaining certain useful characteristics of the pdf  $f(x)$  by investigating the expectations of special cases of  $G(X)$ .

### 4.4.1 Moments of a Distributions

Consider first the case where  $G(X)$  in Eq (4.64) is given as:

$$G(X) = X^k \quad (4.68)$$

for any integer  $k$ . The expectation of this function is known as *the  $k^{th}$  (ordinary) moment of the random variable  $X$*  (or, equivalently, the  $k^{th}$  (ordinary) moment of the pdf,  $f(x)$ ), defined by:

$$m_k = E[X^k] \quad (4.69)$$

#### First (Ordinary) Moment: Mean

Observe that  $m_0 = 1$  always for *all* random variables,  $X$ , and provided that  $E[|X|^k] < \infty$ , then the other  $k$  moments exist; in particular, the first moment

$$m_1 = E(X) = \mu \quad (4.70)$$

Thus, the expected value of  $X$ ,  $E(X)$ , is also the same as the first (ordinary)

moment of  $X$  (or, equivalently, of the pdf  $f(x)$ ).

### Central Moments

Next, consider the case where  $G(X)$  in Eq (4.64) is given as:

$$G(X) = (X - a)^k \quad (4.71)$$

for any constant value  $a$  and integer  $k$ . The expectation of this function is known as *the  $k^{\text{th}}$  moment of the random variable  $X$  about the point  $a$*  (or, equivalently, the  $k^{\text{th}}$  moment of the pdf,  $f(x)$ , about the point  $a$ ). Of particular interest are the moments about the mean value  $\mu$ , defined by:

$$\mu_k = E[(X - \mu)^k] \quad (4.72)$$

known as the *central moments* of the random variable  $X$  (or of the pdf,  $f(x)$ ). Observe from here that  $\mu_0 = 1$ , and  $\mu_1 = 0$ , always, regardless of  $X$  or  $\mu$ ; these therefore provide no particularly useful information regarding the characteristics of any particular  $X$ . However, provided that the conditions of absolute convergence and absolute integrability hold, the higher central moments exist and do in fact provide very useful information about the random variable  $X$  and its distribution.

### Second Central Moment: Variance

Observe from above that the quantity

$$\mu_2 = E[(X - \mu)^2] \quad (4.73)$$

is the lowest central moment of the random variable  $X$  that contains any “meaningful” information about the average deviation of a random variable from its mean value. It is called the *variance* of  $X$  and is sometimes represented as  $\sigma^2(X)$ . Thus,

$$\mu_2 = E[(X - \mu)^2] = \text{Var}(X) = \sigma^2(X). \quad (4.74)$$

Note that

$$\sigma^2(X) = E[(X - \mu)^2] = E(X^2 - 2\mu X + \mu^2) \quad (4.75)$$

so that by the linearity of the  $E[.]$  operator, we obtain:

$$\sigma^2(X) = E(X^2) - \mu^2 = E(X^2) - [E(X)]^2 \quad (4.76)$$

or, in terms of the ordinary moments,

$$\sigma^2(X) = m_2 - \mu^2 \quad (4.77)$$

It is easy to verify the following important properties of  $\text{Var}(X)$ :

1. For constant  $b$ ,

$$\text{Var}(b) = 0 \quad (4.78)$$

2. For constants  $a$  and  $b$ ,

$$\text{Var}(aX + b) = a^2 \text{Var}(X) \quad (4.79)$$

The positive square root of  $\sigma^2$  is called the *standard deviation* of  $X$ , and naturally represented by  $\sigma$ ; it has the same units as  $X$ . The ratio of the standard deviation to the mean value of a random variable, known as the coefficient of variation  $C_v$ , i.e.,

$$C_v = \frac{\sigma}{\mu} \quad (4.80)$$

provides a dimensionless measure of the relative amount of variability displayed by the random variable.

### Third Central Moment: Skewness

The third central moment,

$$\mu_3 = E[(X - \mu)^3] \quad (4.81)$$

is called the *skewness* of the random variable; it provides information about the relative difference that exists between negative and positive deviations from the mean. It is therefore a measure of asymmetry. The dimensionless quantity

$$\gamma_3 = \frac{\mu_3}{\sigma^3} \quad (4.82)$$

known as the *coefficient of skewness*, is often the more commonly used measure precisely because it is dimensionless. For a perfectly symmetric distribution, negative deviations from the mean exactly counterbalance positive deviations, and both  $\gamma_3$  and  $\mu_3$  vanish.

When there are more values of  $X$  to the left of the mean  $\mu$  than to the right, (i.e. when negative deviations from the mean dominate),  $\gamma_3 < 0$  (as is  $\mu_3$ ), and the distribution is said to “skew left” or is “negatively skewed.” Such distributions will have long left tails, as illustrated in Fig 4.3. An example random variable with this characteristic is the gasoline-mileage (in miles per gallon) of cars in the US. While many cars get relatively high gas-mileage, there remains a few classes of cars (SUVs, Hummers, etc) with gas-mileage much worse than the ensemble average. It is this latter class that contribute to the long left tail.

On the other hand, when there are more values of  $X$  to the right of the mean  $\mu$  than to the left, so that positive deviations from the mean dominate, both  $\gamma_3$  and  $\mu_3$  are positive, and the distribution is said to “skew right” or is “positively skewed.” As one would expect, such distributions will have long right tails (see Fig 4.4). An example of this class of random variables is the household income/net-worth in the US. While the vast majority of household incomes/net-worth are moderate, the few truly super-rich whose incomes/net-worth are a few orders of magnitude larger than the ensemble

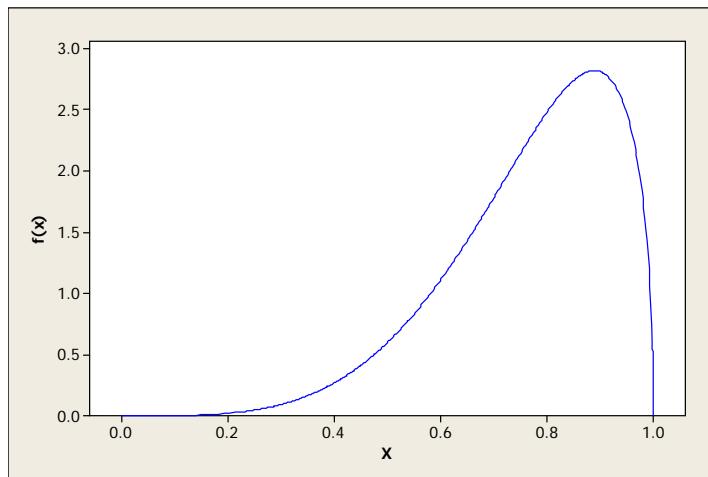


FIGURE 4.3: Distribution of a negatively skewed random variable

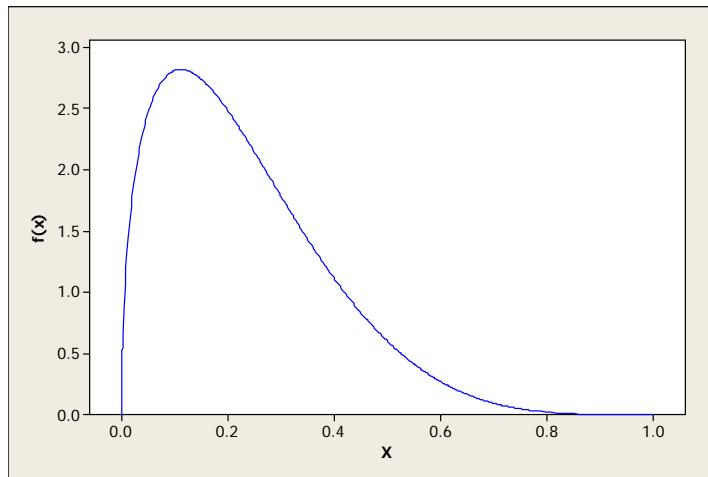
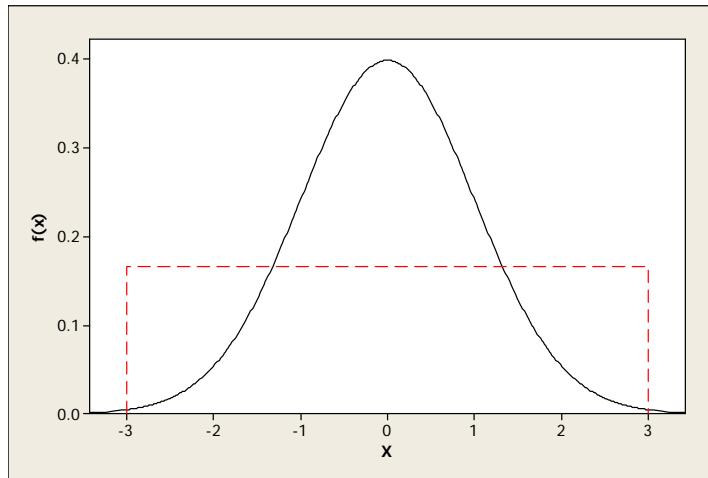


FIGURE 4.4: Distribution of a positively skewed random variable



**FIGURE 4.5:** Distributions with reference kurtosis (solid line) and mild kurtosis (dashed line)

average contribute to the long right tail.

#### Fourth Central Moment: Kurtosis

The fourth central moment,

$$\mu_4 = E[(X - \mu)^4] \quad (4.83)$$

is called the *kurtosis* of the random variable. Sometimes, it is the dimensionless version

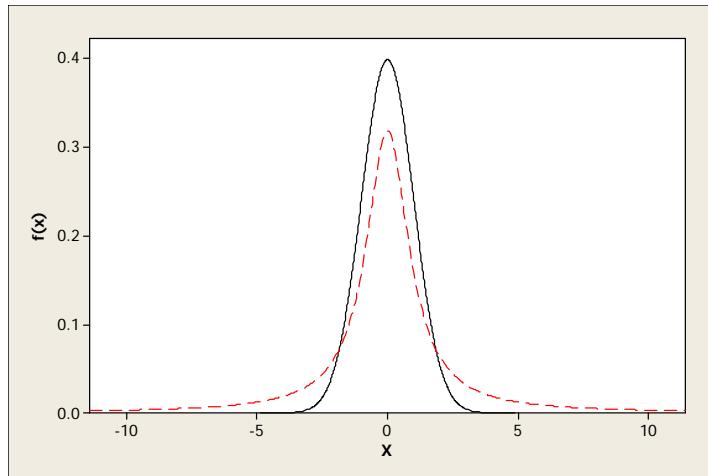
$$\gamma_4 = \frac{\mu_4}{\sigma^4}, \quad (4.84)$$

technically known as the *coefficient of kurtosis*, that is simply called the *kurtosis*. Either quantity is a measure of how peaked or flat a probability distribution is. A high kurtosis random variable has a distribution with a sharper “peak” and thicker “tails;” the low kurtosis random variable on the other hand has a distribution with a more rounded, flatter peak, with broader “shoulders.”

For reasons discussed later, the value  $\gamma_4 = 3$  is the accepted “normal” reference for kurtosis, so that distributions for which  $\gamma_4 < 3$  are said to be *platykurtic* (mildly peaked) while those for which  $\gamma_4 > 3$  are said to be *leptokurtic* (sharply peaked). Figures 4.5 and 4.6 show a reference distribution with kurtosis  $\gamma_4 = 3$ , in the solid lines, compared to a distribution with *mild kurtosis* (actually  $\gamma_4 = 1.8$ ) (dashed line in Fig 4.5), and a distribution with *high kurtosis* (dashed line in Fig 4.6).

#### Practical Applications

Of course, it is possible to compute as many moments (ordinary or central) of



**FIGURE 4.6:** Distributions with reference kurtosis (solid line) and high kurtosis (dashed line)

a distribution as we wish — and we shall shortly present a general expression from which one can generate all such moments; but the four specifically singled out above have been the most useful for characterizing random variables and their distributions, in practice. They tell us much about the random variable we are dealing with.

The first (ordinary) moment,  $m_1$  or  $\mu$ , tells us about the location of the “center of gravity” (centroid) of the random variable, its mean value; and, as we show later, it is a popular candidate for the single value *most representative* of the ensemble. The second central moment,  $\mu_2$  or  $\sigma^2$ , the variance, tells us how tightly clustered — or broadly dispersed — the random variable is around its mean. The third central moment,  $\mu_3$ , the skewness, tells us whether lower extreme values of the random variable are farther to the left of the centroid (the ensemble average) than the higher extreme values are to the right (as is the case with automobile gas-mileage in the US), or vice versa, with higher extreme values significantly farther to the right of the centroid than the lower extreme values (as is the case with household incomes/net worth in the US).

Just like the third central moment tells us how much of the average *deviation from the mean* is due to infrequent extreme values, the fourth central moment,  $\mu_4$  (the kurtosis) tells us how much of the *variance* is due to infrequent extreme deviations. With sharper peaks and thicker tails, extreme values in the tails contribute more to the variance, and the kurtosis is high (as in Fig 4.6); with flatter peaks and very little in terms of tails (as in Fig 4.5), there will be more contributions to the variance from central values, which naturally show modest deviations from the mean, and very little contribution from the extreme values; the kurtosis will therefore be lower.

Finally, we note that moments of a random variable are not merely interesting theoretical characteristics; they have significant practical applications. For example, polymers, being macromolecules with non-uniform molecular weights (because random events occurring during the manufacturing process ensure that polymer molecules grow to varying sizes) are primarily characterized by their molecular weight distributions (MWDs). Not surprisingly, therefore, the performance of a polymeric material depends critically on its MWD: for instance, with most elastomers, a narrow distribution (very low second central moments) is associated with poor processing but superior mechanical properties.

MWDs are so important in polymer chemistry and engineering that a wide variety of analytical techniques have been developed for experimental determination of the MWD and the following special molecular weight averages that are in common use:

1.  $M_n$ , the number average molecular weight, is the ratio of the first (ordinary) moment to the zeroth ordinary moment. (In polymer applications, the MWD, unlike a pdf  $f(x)$ , is *not* normalized to sum or integrate to 1. The zeroth moment of the MWD is therefore not 1; it is the total number of molecules present in the sample of interest.)
2.  $M_w$ , the weight average molecular weight, is the ratio of the second moment to the first moment; and
3.  $M_z$ , the so-called  $z$  average molecular weight, is the ratio of the third moment to the second.

One other important practical characteristic of the polymeric material is its “polydispersity index”, PDI, the ratio of  $M_w$  to  $M_n$ . A measure of the breadth of the MWD, it is always  $> 1$  and approximately 2 for most linear polymers; for highly branched polymers, it can be as high as 20 or even higher.

What is true of polymers is also true of particulate products such as granulated sugar, or fertilizer granules sold in bags. These products are made up of particles with non-uniform sizes and are characterized by their particle size distributions. The behavior of these products, whether it is their flow characteristics, or how they dissolve in solution, are determined by the moments of these distributions.

#### 4.4.2 Moment Generating Function

When  $G(X)$  in Eq (4.64) is given as:

$$G(X) = e^{tX} \quad (4.85)$$

the expectation of this function, when it exists, is the function:

$$M_X(t) = \begin{cases} \sum_i e^{tx_i} f(x_i); & \text{for discrete } X; \\ \int_{-\infty}^{\infty} e^{tx} f(x) dx; & \text{for continuous } X \end{cases} \quad (4.86)$$

a function of the real-valued variable,  $t$ , known as the *moment generating function* (MGF) of  $X$ .  $M_X(t)$  is so called because all the (ordinary) moments of  $X$  can be generated from it as follows:

By definition,

$$M_X(t) = E(e^{tX}) \quad (4.87)$$

and by differentiating with respect to  $t$ , we obtain,

$$\begin{aligned} M'_X(t) &= \frac{d}{dt} E(e^{tX}) = E\left[\frac{d}{dt}(e^{tX})\right] \\ &= E(Xe^{tX}) \end{aligned} \quad (4.88)$$

(The indicated swapping of the order of the differentiation and expectation operators is allowed under conditions that essentially imply the existence of the moments.) From here we easily obtain, for  $t = 0$ , that:

$$M'_X(0) = E(X) = m_1 \quad (4.89)$$

the first (ordinary) moment. Similarly, by differentiating once more, we obtain:

$$M''_X(t) = \frac{d}{dt} E(Xe^{tX}) = E(X^2 e^{tX}) \quad (4.90)$$

so that, for  $t = 0$ ,

$$M''_X(0) = E(X^2) = m_2 \quad (4.91)$$

and in general, after  $n$  such differentiations, we obtain

$$M_X^{(n)}(0) = E[X^n] = m_n \quad (4.92)$$

Now, it is also possible to establish this result by considering the following Taylor series expansion about the point  $t = 0$ ,

$$e^{tX} = 1 + Xt + \frac{X^2}{2}t^2 + \frac{X^3}{3!}t^3 + \dots \quad (4.93)$$

Clearly, this infinite series converges only under certain conditions. For those random variables,  $X$ , for which the series does not converge,  $M_X(t)$  does not exist; but when it exists, this series converges, and by repeated differentiation of Eq (4.93) with respect to  $t$ , followed by taking expectations, we are then able to establish the result in Eq (4.92).

The following are some important properties of the MGF.

1. *Uniqueness:* The MGF,  $M_X(t)$ , does not exist for all random variables,  $X$ ; but when it exists, it uniquely determines the distribution, so that if two random variables have the same MGF, they have the same distribution. Conversely, random variables with different MGF's have different distributions.

2. *Linear Transformations:* If two random variables  $Y$  and  $X$  are related according to the linear expression:

$$Y = aX + b \quad (4.94)$$

for constant  $a$  and  $b$ , then:

$$M_Y(t) = e^{bt} M_X(at) \quad (4.95)$$

3. *Independent Sums:* For independent random variables  $X$  and  $Y$  with respective MGF's  $M_X(t)$ , and  $M_Y(t)$ , the MGF of their sum  $Z = X + Y$  is:

$$M_Z(t) = M_{X+Y}(t) = M_X(t)M_Y(t) \quad (4.96)$$

#### **Example 4.7 MOMENT GENERATING FUNCTION OF A CONTINUOUS RANDOM VARIABLE**

Find the MGF  $M_X(t)$ , for the random variable,  $X$ , the residence time in a CSTR, whose pdf is given in Eq (4.41).

##### **Solution:**

In this case, the required  $M_X(t)$  is given by:

$$M_X(t) = E(e^{tX}) = \frac{1}{\tau} \int_0^\infty e^{tx} e^{-x/\tau} dx = \frac{1}{\tau} \int_0^\infty e^{-\frac{(1-\tau t)x}{\tau}} dx \quad (4.97)$$

Upon integrating the RHS appropriately, we obtain,

$$M_X(t) = - \left( \frac{1}{1 - \tau t} \right) e^{-\frac{(1-\tau t)x}{\tau}} \Big|_0^\infty \quad (4.98)$$

$$= \frac{1}{1 - \tau t} \quad (4.99)$$

From here, one easily obtains:  $m_1 = \tau$ ;  $m_2 = \tau^2$ ,  $\dots$ ,  $m_k = \tau^k$ .

#### **4.4.3 Characteristic Function**

As alluded to above, the MGF does not exist for all random variables, a fact that sometimes limits its usefulness. However, a similarly defined function, the *characteristic function*, shares all the properties of the MGF but does not suffer from this primary limitation: it exists for all random variables.

When  $G(X)$  in Eq (4.64) is given as:

$$G(X) = e^{j t X} \quad (4.100)$$

where  $j$  is the complex variable  $\sqrt{(-1)}$ , then the function of the real-valued variable  $t$  defined as,

$$\varphi_X(t) = E(e^{j t X}) \quad (4.101)$$

i.e.

$$\varphi_X(t) = \begin{cases} \sum_i e^{jtx_i} f(x_i); & \text{for discrete } X; \\ \int_{-\infty}^{\infty} e^{jtx} f(x) dx; & \text{for continuous } X \end{cases} \quad (4.102)$$

is known as the *characteristic function* (CF) of the random variable  $X$ .

Because of the definition of the complex exponential, whereby

$$e^{jtx} = \cos(tx) + j \sin(tx) \quad (4.103)$$

observe that

$$|e^{jtx}| = \cos^2(tx) + \sin^2(tx) = 1 \quad (4.104)$$

so that  $E(|e^{jtx}|) = 1 < \infty$ , always, regardless of  $X$ , with the direct implication that  $\varphi_X(t) = E(e^{jtx})$  always exists for all random variables. Thus, anything one would have typically used the MGF for (e.g., for deriving limit theorems in advanced courses in probability), one can always substitute the CF when the MGF does not exist.

The reader familiar with Laplace transforms and Fourier transforms will probably have noticed the similarities between the former and the MGF (see Eq (4.86)), and between the latter and the CF (see Eq (4.102)). Furthermore, the relationship between these two probability functions are also reminiscent of the relationship between the two transforms: not all functions have Laplace transforms; the Fourier transform, on the other hand, does not suffer such limitations.

We now state, without proof, that given the expression for the characteristic function in Eq (4.102), there is a corresponding *inversion formula* whereby  $f(x)$  is recovered from  $\varphi_X(t)$ , given as follows:

$$f(x) = \begin{cases} \lim_{b \rightarrow \infty} \frac{1}{2b} \int_{-b}^b e^{-jtx} \varphi_X(t) dt; & \text{for discrete } X; \\ \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-jtx} \varphi_X(t) dt; & \text{for continuous } X \end{cases} \quad (4.105)$$

In fact, the two sets of equations, Eqs (4.102) and (4.105), are formal Fourier transform pairs precisely as in other engineering applications of the theory of Fourier transforms. These transform pairs are extremely useful in obtaining the pdfs of functions of random variables, most especially sums of random variables. As with classic engineering applications of the Fourier (and Laplace) transform, the characteristic functions of the functions of independent random variables in question are obtained first, being easier to obtain directly than the pdfs; the inversion formula is subsequently invoked to recover the desired pdfs. This strategy is employed at appropriate places in upcoming chapters.

#### 4.4.4 Additional Distributional Characteristics

Apart from the mean, variance and other higher moments noted above, there are other characteristic attributes of importance.

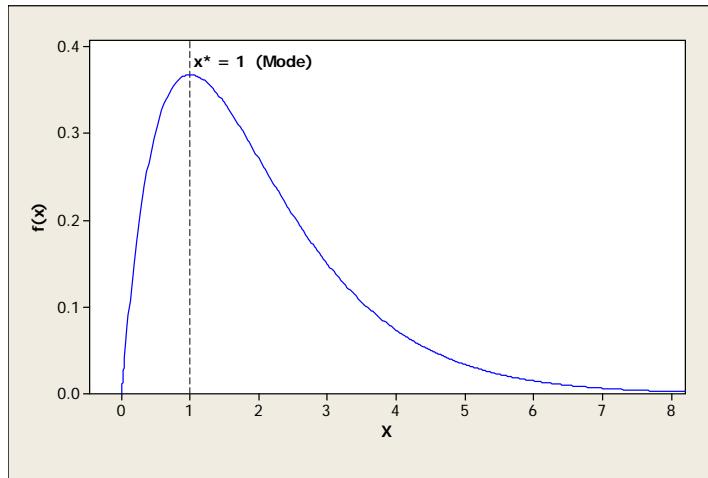


FIGURE 4.7: The pdf of a continuous random variable  $X$  with a mode at  $x = 1$

### Mode

The mode,  $x^*$ , of a distribution is that value of the random variable for which the pdf achieves a (local) maximum. For a discrete random variable, it is the value of  $X$  that possesses the maximum probability (the most “popular” value); i.e.

$$\arg \max_x \{P(X = x)\} = x^* \quad (4.106)$$

For a continuous random variable with a differentiable pdf, it is the value of  $x$  for which

$$\frac{df(x)}{dx} = 0; \frac{d^2 f(x)}{dx^2} < 0 \quad (4.107)$$

as shown in Fig 12.21. A pdf having only one such maximum value is said to be *unimodal*; if more than one such maximum value exists, the distribution is said to be *multimodal*.

### Median

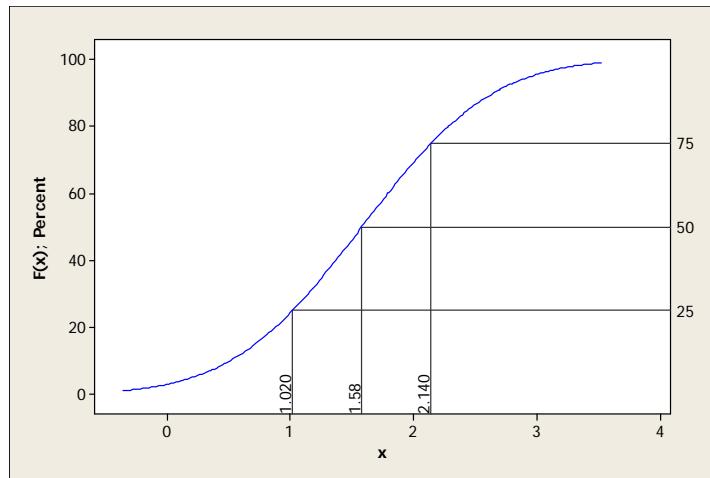
The median of a distribution is that mid-point value  $x_m$  for which the cumulative distribution is exactly  $1/2$ , i.e.

$$F(x_m) = P(X < x_m) = P[X > x_m] = 0.5 \quad (4.108)$$

For a continuous random variable,  $x_m$  is the value for which

$$\int_{-\infty}^{x_m} f(x)dx = \int_{x_m}^{\infty} f(x)dx = 0.5 \quad (4.109)$$

(For the discrete random variable, replace the integral above with appropriate



**FIGURE 4.8:** The cdf of a continuous random variable  $X$  showing the lower and upper quartiles and the median

sums.) Observe therefore that the median,  $x_m$ , divides the total range of the random variable into two parts with equal probability.

For a symmetric unimodal distribution, the mean, mode and median coincide; they are different for asymmetric (skewed) distributions.

### Quartiles

The concept of a median, which divides the cdf at the 50% point, can be extended to other values indicative of other fractional sectioning off of the cdf. Thus, by referring to the median as  $x_{0.5}$ , or  $x_{50}$ , we are able to define, in the same spirit, the following values of the random variable,  $x_{0.25}$  and  $x_{0.75}$  (or, in terms of percentages,  $x_{25}$  and  $x_{75}$  respectively) as follows:

$$F(x_{0.25}) = 0.25 \quad (4.110)$$

that value of  $X$  below which a quarter of the population resides; and

$$F(x_{0.75}) = 0.75 \quad (4.111)$$

the value of  $X$  below which lies three quarters of the population. These values are known respectively as the *lower* and *upper quartiles* of the distribution because, along with the median  $x_{0.5}$ , these values divide the population into four quarters, each part with equal probability.

These concepts are illustrated in Fig 4.8 where the lower quartile is located at  $x = 1.02$ ; the median at  $x = 1.58$  and the upper quartile at  $x = 2.14$ . Thus, for this particular example,  $P(X < 1.02) = 0.25$ ;  $P(1.02 < X < 1.58) = 0.25$ ;  $P(1.58 < X < 2.14) = 0.25$  and  $P(X > 2.14) = 0.25$ .

There is nothing restricting us to dividing the population in halves (median) or in quarters (quartiles); in general, for any  $0 < q < 1$ , the  $q^{\text{th}}$  quantile is defined as that value  $x_q$  of the random variable for which

$$F(x_q) = \int_{-\infty}^{x_q} f(x)dx = q \quad (4.112)$$

for a continuous random variable (with the integral replaced by the appropriate sum for the discrete random variable).

This quantity is sometimes defined instead in terms of *percentiles*, in which case, the  $q^{\text{th}}$  quantile is simply the  $100q$  percentile. Thus, the median is equivalently the “half quantile”, the “50th percentile”, or the “second quartile.”

#### 4.4.5 Entropy

A concept to be explored more completely in Chapter 10 is concerned with quantifying the “information content” contained in the statement, “ $X = x$ ”, i.e. that the (discrete) random variable  $X$  has been observed to take on the specific value  $x$ . Whatever this information content is, it will clearly be related to the pdf,  $f(x)$ ; in fact, it has been shown to be defined as:

$$I[f(x)] = -\log_2 f(x) \quad (4.113)$$

Now, when  $G(X)$  in Eq (4.64) is defined as:

$$G(X) = -\log_2 f(x) \quad (4.114)$$

then the expectation in this case is the function  $\mathcal{H}(x)$ , defined as:

$$\mathcal{H}(x) = \begin{cases} -\sum_i f(x_i) \log_2 f(x_i); & \text{for discrete } X \\ -\int_{-\infty}^{\infty} f(x) \log_2 f(x) dx; & \text{for continuous } X \end{cases} \quad (4.115)$$

known as the *entropy* of the random variable, or, its *mean information content*. Chapter 10 explores how to use the concept of information and entropy to develop appropriate probability models for practical problems in science and engineering.

#### 4.4.6 Probability Bounds

We now know that the pdf  $f(x)$  of a random variable contains all the information about it to enable us compute the probabilities of occurrence of various outcomes of interest. As valuable as this is, there are times when all we need are bounds on probabilities, not exact values. We now discuss some of the most important results regarding bounds on probabilities that can be determined for any general random variable,  $X$  without specific reference to

any particular pdf. These results are very useful in analyzing the behavior of random phenomena and have practical implications in determining values of unknown population parameters.

We begin with a general lemma from which we then derive two important results.

**Lemma:** Given a random variable  $X$  (with a pdf  $f(x)$ ), and  $G(X)$  a function of this random variable such that  $G(X) > 0$ , for an arbitrary constant,  $c > 0$ ,

$$P(G(X) \geq c) \leq \frac{E[G(X)]}{c} \quad (4.116)$$

There are several different ways of proving this result; one of the most direct is shown below.

**Proof:** By definition,

$$E[G(X)] = \int_{-\infty}^{\infty} G(x)f(x)dx \quad (4.117)$$

If we now divide the real line  $-\infty < x < \infty$  into two mutually exclusive regions,  $A = \{x : G(x) \geq c\}$  and  $B = \{x : G(x) < c\}$ , i.e.  $A$  is that region on the real line where  $G(x) \geq c$ , and  $B$  is what is left, then, Eq (4.117) becomes:

$$E[G(X)] = \int_A G(x)f(x)dx + \int_B G(x)f(x)dx \quad (4.118)$$

and since  $G(X)$  is non-negative, the second integral is  $\geq 0$ , so that

$$E[G(X)] \geq \int_A G(x)f(x)dx \geq \int_A cf(x)dx \quad (4.119)$$

where the last inequality arises because, for all  $x \in A$ , (the region over which we are integrating)  $G(x) \geq c$ , with the net results that:

$$E[G(X)] \geq cP(G(X) \geq c) \quad (4.120)$$

because the last integral is, by definition,  $cP(A)$ . From here, we now obtain

$$P[G(X) \geq c] \leq \frac{E[G(X)]}{c} \quad (4.121)$$

as required.

This remarkable result holds for all random variables,  $X$ , and for *any* non-negative functions of the random variable,  $G(X)$ . Two specific cases of  $G(X)$  give rise to results of special interest.

### Markov's Inequality

When  $G(X) = X$ , Eq (4.116) immediately becomes:

$$P(X \geq c) \leq \frac{E(X)}{c} \quad (4.122)$$

a result known as *Markov's inequality*. It allows us to place bounds on probabilities when only the mean value of a random variable is known. For example, if the average number of inclusions on glass sheets manufactured in a specific site is known to be 2, then according to Markov's inequality, the probability of finding a glass sheet containing 5 or more inclusions at this manufacturing site *can never exceed*  $2/5$ . Thus if glass sheets containing 5 or more inclusions are considered unsaleable, without reference to any specific probability model of the random phenomenon in question, the plant manager concerned about making unsaleable product can, by appealing to Markov's inequality, be sure that things will never be worse than 2 in 5 unsaleable products.

It is truly remarkable, of course, that such statements can be made at all; but in fact, this inequality is actually quite conservative. As one would expect, with an appropriate probability model, one can be even more precise. (Table 2.1 in Chapter 2 in fact shows that the actual probability of obtaining 5 or more inclusions on glass sheets manufactured at this site is 0.053, nowhere close to the upper limit of 0.4 given by Markov's inequality.)

### Chebychev's Inequality

Now let  $G(X) = (X - \mu)^2$ , and  $c = k^2\sigma^2$ , where  $\mu$  is the mean value of  $X$ , and  $\sigma^2$  is the variance, i.e.  $\sigma^2 = E[(x - \mu)^2]$ . In this case, Eq (4.116) becomes

$$P[(X - \mu)^2 \geq k^2\sigma^2] \leq \frac{1}{k^2} \quad (4.123)$$

which may be simplified to:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}, \quad (4.124)$$

a result known as Chebychev's inequality. The implication is that  $1/k^2$  is an upper bound for the probability that any random variable will take on values that deviate from the mean by more than  $k$  standard deviations. This is still a rather weak inequality in the sense that in most cases, the indicated probability is far less than  $1/k^2$ . Nevertheless, the added information of known  $\sigma$  helps sharpen the bounds a bit, when compared to Markov's inequality. For example, if we now add to the glass sheets inclusions information the fact that the variance is 2 (so that  $\sigma = \sqrt{2}$ ), then, the desired probability  $P(X \geq 5)$

now translates to  $P(|X - 2| \geq 3)$  since  $\mu = 2$ . In this case, therefore,  $k\sigma = 3$ , and from Chebychev's inequality, we obtain:

$$P(|X - 2| \geq 3) \leq \frac{\sigma^2}{9} = \frac{2}{9} \quad (4.125)$$

an upper bound which, even though still conservative, is nevertheless much sharper than the  $2/5$  obtained earlier from Markov's inequality.

Chebychev's inequality plays a significant role in Chapter 8 in establishing a fundamental result relating relative frequencies in repeatable experiments to the probabilities of occurrence of events.

## 4.5 Special Derived Probability Functions

In studying phenomena involving lifetimes (of humans and other living organisms, or equipment, or, for that matter, social movements), or more generally in studying the elapsed time until the occurrence of specific events — studies that encompass the related problem of reliability of equipment and systems — the application of probability theory obviously still involves the use of the pdf  $f(x)$  and the cdf  $F(x)$ , but in specialized forms unique to such problems. The following is a discussion of special probability functions, derived from  $f(x)$  and  $F(x)$ , that have been customized for such applications. As a result, these special probability functions are exclusively for random variables that are (a) continuous, and (b) non-negative; they do not exist for random variables that do not satisfy these conditions.

### 4.5.1 Survival Function

The survival function,  $S(x)$ , is the probability that the random variable  $X$  exceeds the specific value  $x$ ; in lifetime applications, this translates to the probability that the object of study “survives” beyond the value  $x$ , i.e.

$$S(x) = P(X > x) \quad (4.126)$$

From the definition of the cdf,  $F(x)$ , we see immediately that

$$S(x) = 1 - F(x) \quad (4.127)$$

so that where  $F(x)$  is a monotonically increasing function of  $x$  that starts at 0 and ends at 1,  $S(x)$  is the exact mirror image, monotonically decreasing from 1 to 0.

#### Example 4.8 SURVIVAL FUNCTION OF A CONTINUOUS RANDOM VARIABLE

Find the survival function  $S(x)$ , for the random variable,  $X$ , the residence time in a CSTR, whose pdf is given in Eq (4.41). This function directly provides the probability that any particular dye molecule “survives” in the CSTR beyond a time  $x$ .

**Solution:**

Observe first that this random variable is continuous and non-negative so that the desired  $S(x)$  does in fact exist. The required  $S(x)$  is given by

$$S(x) = \frac{1}{\tau} \int_x^{\infty} e^{-x/\tau} dx = e^{-x/\tau} \quad (4.128)$$

We could equally well have arrived at the result by noting that the cdf  $F(x)$  for this random variable is given by:

$$F(x) = (1 - e^{-x/\tau}). \quad (4.129)$$

Note from Eq (4.128) that with increasing  $x$  (residence time), survival becomes smaller; i.e. the probability of still finding a dye molecule in the reactor after a time  $x$  has elapsed diminishes exponentially with  $x$ .

#### 4.5.2 Hazard Function

In reliability and life-testing studies, it is useful to have a means of directly computing the probability of failure in the intervals *beyond the current time*,  $x$ , for entities that have survived thus far; i.e. probabilities of failure conditioned on survival until  $x$ . The hazard function,  $h(x)$ , defined as follows:

$$h(x) = \frac{f(x)}{S(x)} = \frac{f(x)}{1 - F(x)} \quad (4.130)$$

provides just such a function. It does for *future* failure what  $f(x)$  does for lifetimes in general. Recall that by definition, because  $X$  is continuous,  $f(x)$  provides the (unconditional) probability of a lifetime in the infinitesimal interval  $\{x_i < X < x_i + dx\}$  as  $f(x_i)dx$ ; in the same manner, the probability of failure occurring in that same interval, given that the object of study survived until the beginning of the current time interval,  $x_i$ , is given by  $h(x_i)dx$ . In general

$$h(x)dx = \frac{f(x)dx}{S(x)} = \frac{P(x < X < x + dx)}{P(X > x)} \quad (4.131)$$

so that, from the definition of conditional probability given in Chapter 3,  $h(x)dx$  is seen as equivalent to  $P(x < X < x + dx|X > x)$ .  $h(x)$  is therefore sometimes referred to as the “death rate” of “failure rate” at  $x$  of those surviving until  $x$  (i.e. of those at “risk” at  $x$ ); it describes how the risk of failure changes with age.

#### Example 4.9 HAZARD FUNCTION OF A CONTINUOUS RANDOM VARIABLE

Find the hazard function  $h(x)$ , for the random variable,  $X$ , the residence time in a CSTR.

**Solution:**

From the given pdf and the survival function obtained in Example 4.8 above, the required function  $h(x)$  is given by,

$$h(x) = \frac{\frac{1}{\tau}e^{-x/\tau}}{e^{-x/\tau}} = \frac{1}{\tau} \quad (4.132)$$

a constant, with the interesting implication that the probability that a dye molecule exits the reactor immediately after time  $x$ , given that it had stayed in the reactor until then, is independent of  $x$ . Thus molecules that have survived in the reactor until  $x$  have the same chance of exiting the reactor immediately after this time as the chance of exiting at any other time in the future — no more, no less. Such a random variable is said to be “memoryless;” how long it lasts beyond the current time does not depend on its current age.

#### 4.5.3 Cumulative Hazard Function

Analogous to the cdf,  $F(x)$ , the cumulative hazard function,  $H(x)$ , is defined as:

$$H(x) = \int_0^x h(u)du \quad (4.133)$$

It can be shown that  $H(x)$  is related to the more well-known  $F(x)$  according to

$$F(x) = 1 - e^{-H(x)} \quad (4.134)$$

and that the relationship between  $S(x)$  and  $H(x)$  is given by:

$$S(x) = e^{-H(x)} \quad (4.135)$$

or, conversely,

$$H(x) = -\log[S(x)] \quad (4.136)$$

#### 4.6 Summary and Conclusions

We are now in a position to look back at this chapter and observe, with some perspective, how the introduction of the seemingly innocuous random variable,  $X$ , has profoundly affected the analysis of randomly varying phenomena in a manner analogous to how the introduction of the “unknown quantity,”  $x$ , transformed algebra and the solution of algebraic problems. We have seen how the random variable,  $X$ , maps the sometimes awkward and

tedious sample space,  $\Omega$ , into a space of real numbers; how this in turn leads to the emergence of  $f(x)$ , the probability distribution function (pdf); and how  $f(x)$  has essentially supplanted and replaced the probability set function,  $P(A)$ , the probability analysis tool in place at the end of Chapter 3.

The full significance of the role of  $f(x)$  in random phenomena analysis may not be completely obvious now, but it will become more so as we progress in our studies. So far, we have used it to characterize the random variable in terms of its mathematical expectation, and the expectation of various other functions of the random variable. And this has led, among other things, to our first encounter with the *mean*, *variance*, *skewness* and *kurtosis*, of a random variable, important descriptors of data that we are sure to encounter again later (in Chapter 12 and beyond).

Despite initial appearances, every single topic discussed in this chapter finds useful application in later chapters. In the meantime, we have taken pains to try and breathe some “practical life” into many of these typically dry and formal definitions and mathematical functions. But if some, especially the moment generating function, the characteristic function, and entropy, still appear to be of dubious practical consequence, such lingering doubts will be dispelled completely by Chapters 6, 8, 9 and 10. Similarly, the probability bounds (especially Chebyshev’s inequality) will be employed in Chapter 8, and the special functions of Section 4.5 will be used extensively in their more natural setting in Chapter 23.

The task of building an efficient machinery for random phenomena analysis, which began in Chapter 3, is now almost complete. But before the generic pdf,  $f(x)$ , introduced and characterized in this chapter begins to take on specific, distinct personalities for various random phenomena, some residual issues remain to be addressed in order to complete the development of the probability machinery. Specifically, the discussion in this chapter will be extended to higher dimensions in Chapter 5, and the characteristics of functions of random variables will be explored in Chapter 6. Chapter 7 is devoted to two application case studies that put the complete set of discussions in Part II in perspective.

Here are some of the main points of the chapter again.

- *Formally*, the random variable,  $X$ —discrete or continuous—assigns to each element  $\omega \in \Omega$ , one and only one real number,  $X(\omega) = x$ , thereby mapping  $\Omega$  onto a new space,  $V$ ; informally it is an experimental outcome whose numerical value is subject to random variations with each exact replicate trial of the experiment.
- The introduction of the random variable,  $X$ , leads directly to the emergence of  $f(x)$ , the probability distribution function; it represents how the probabilities of occurrence of all the possible outcomes of the random experiment of interest are distributed over the entire random variable space, and is a direct extension of  $P(A)$ .

- The cumulative distribution function (cdf),  $F(x)$ , is  $P(X \leq x)$ ; if discrete  $F(x_i) = \sum_{j=0}^i f(x_j)$ ; if continuous,  $\int_{-\infty}^{x_i} f(x)dx$ , so that if differentiable,  $\frac{dF(x)}{dx} = f(x)$ .
- The mathematical expectation of a random variable,  $E(X)$ , is defined as;

$$E(X) = \begin{cases} \sum_i x_i f(x_i); & \text{discrete;} \\ \int_{-\infty}^{\infty} x f(x) dx; & \text{continuous} \end{cases}$$

It exists only when  $\sum_i |x_i|f(x_i) < \infty$  (absolute convergence for discrete random variables) or  $\int_{-\infty}^{\infty} |x|f(x) dx < \infty$  (absolute integrability for continuous random variables).

- $E[G(X)]$  provides various characterizations of the random variables,  $X$ , for various functions  $G(X)$ :
  - $G(X) = (X - \mu)^k$  yields the  $k^{th}$  moment of  $X$ ;
  - $G(X) = e^{tX}$  and  $G(X) = e^{jtX}$  respectively yield the moment generating function (MGF), and the characteristic function (CF), of  $X$ ;
  - $G(X) = -\log_2 f(x)$  yields the entropy of  $X$ .
- The *mean*, indicates the central location or “center of gravity” of the random variable while the *variance*, *skewness* and *kurtosis* indicate the shape of the distribution in relation to the mean. Additional characterization is provided by the *mode*, where the distribution is maximum and by the *median*, which divides the distribution into two equal probability halves; the *quartiles*, which divide the distribution into four equal probability quarters, or more generally, the percentiles, which divide the distribution into 100 equal probability portions.
- Lifetimes and related phenomena are more conveniently studied with special probability functions, which include:
  - The survival function,  $S(x)$ , the probability that  $X$  exceeds the value  $x$ ; by definition, it is related to  $F(x)$  according to  $S(x) = 1 - F(x)$ ;
  - The hazard function,  $h(x)$ , which does for future failure probabilities what  $f(x)$  does for lifetime probabilities; and
  - The cumulative hazard function,  $H(x)$ , which is to the hazard function,  $h(x)$ , what the cdf  $F(x)$  is to the pdf  $f(x)$ .

## REVIEW QUESTIONS

1. Why is the raw sample space,  $\Omega$ , often tedious to describe and inefficient to analyze mathematically?
2. Through what means is the general sample space converted into a space with real numbers?
3. Formally, what is a random variable?
4. What two mathematical transformations occur as a consequence of the formal introduction of the random variable,  $X$ ?
5. How is the induced probability set function,  $P_X$ , related to the probability set function,  $P$ , defined on  $\Omega$ ?
6. What is the pre-image,  $\Gamma_A$ , of the set  $A$ ?
7. What is the relationship between the random variable,  $X$ , and the associated real number,  $x$ ? What does the expression,  $P(X = x)$  indicate?
8. When does the sample space,  $\Omega$ , naturally occur in the form of the random variable space,  $V$ ?
9. Informally, what is a random variable?
10. What is the difference between a discrete random variable and a continuous one?
11. What is the pdf,  $f(x)$ , and what does it represent for the random variable,  $X$ ?
12. What is the relationship between the pdf,  $f(x_i)$ , and the cdf,  $F(x_i)$ , for a discrete random variable,  $X$ ?
13. What is the relationship between the pdf,  $f(x)$ , and the cdf,  $F(x)$ , for a continuous random variable,  $X$ ?
14. Define mathematically the expected value,  $E(X)$ , for a discrete random variable and for a continuous one.
15. What conditions must be satisfied for  $E(X)$  to exist?
16. Is  $E(X)$  a random variable and does it have units?
17. What is the relationship between the expected value,  $E(X)$ , and the mean value,  $\mu$  of a random variable (or equivalently, of its distribution)?
18. Distinguish between ordinary moments and central moments of a random variable.

- 19.** What are the common names by which the second, third and fourth central moments of a random variable are known?
- 20.** What is  $C_v$ , the coefficient of variation of a random variable?
- 21.** What is the distinguishing characteristic of a skewed distribution (positive or negative)?
- 22.** Give an example each of a negatively skewed and a positively skewed randomly varying phenomenon.
- 23.** What do the mean, variance, skewness, and kurtosis tell us about the distribution of the random variable in question?
- 24.** What do  $M_n$ ,  $M_w$ , and  $M_z$  represent for a polymer material?
- 25.** What is the polydispersity index of a polymer and what does it indicate about the molecular weight distribution?
- 26.** Define the moment generating function (MGF) of a random variable,  $X$ . Why is it called by this name?
- 27.** What is the uniqueness property of the MGF?
- 28.** Define the characteristic function of a random variable,  $X$ . What distinguishes it from the MGF?
- 29.** How are the MGF and characteristic function (CF) of a random variable related to the Laplace and Fourier transforms?
- 30.** Define the mode, median, quartiles and percentiles of a random variable.
- 31.** Within the context of this chapter, what is “Entropy”?
- 32.** Define Markov’s inequality. It allows us to place probability bounds when what is known about the random variable?
- 33.** Define Chebychev’s inequality.
- 34.** Which probability bound is sharper, the one provided by Markov’s inequality or the one provided by Chebychev’s?
- 35.** What are the defining characteristics of those random variables for which the special probability functions, the survival and hazard functions, are applicable? These functions are used predominantly in studying what types of phenomena?
- 36.** Define the survival function,  $S(x)$ . How is it related to the cdf,  $F(x)$ ?

- 37.** Define the hazard function,  $h(x)$ . How is it related to the pdf,  $f(x)$ ?
- 38.** Define the cumulative hazard function,  $H(x)$ . How is it related to the cdf,  $F(x)$ , and the survival function,  $S(x)$ ?

## EXERCISES

### Section 4.1

**4.1** Consider a family that plans to have a total of three children; assuming that they will not have any twins, generate the sample space,  $\Omega$ , for the possible outcomes. By defining the random variable,  $X$  as the total number of female children born to this family, obtain the corresponding random variable space,  $V$ . Given that this particular family is genetically predisposed to having boys, with a probability,  $p = 0.75$  of giving birth to a boy, obtain the probability that this family will have three boys and compare it to the probability of having other combinations.

**4.2** Revisit Example 4.1 in the text, and this time, instead of tossing a coin three times, it is tossed 4 times. Generate the sample space,  $\Omega$ ; and using the same definition of  $X$  as the total number of tails, obtain the random variable space,  $V$ , and compute anew the probability of  $A$ , the event that  $X = 2$ .

**4.3** Given the spaces  $\Omega$  and  $V$  for the double dice toss experiment in Example 4.3 in the text,

- (i) Compute the probability of the event  $A$  that  $X = 7$ ;
- (ii) If  $B$  is the event that  $X = 6$ , and  $C$  the event that  $X = 10$  or  $X = 11$ , compute  $P(B)$  and  $P(C)$ .

### Section 4.2

**4.4** Revisit Example 4.3 in the text on the double dice toss experiment and obtain the complete pdf  $f(x)$  for the entire random variable space. Also obtain the cdf,  $F(x)$ . Plot both distribution functions.

**4.5** Given the following probability distribution function for a discrete random variable,  $X$ ,

| $x$    | 1    | 2    | 3    | 4    | 5    |
|--------|------|------|------|------|------|
| $f(x)$ | 0.10 | 0.25 | 0.30 | 0.25 | 0.10 |

- (i) Obtain the cdf  $F(x)$ .
- (ii) Obtain  $P(X \leq 3)$ ;  $P(X < 3)$ ;  $P(X > 3)$ ;  $P(2 \leq X \leq 4)$

**4.6** A particular *discrete* random variable,  $X$ , has the cdf

$$F(x) = \left(\frac{x}{n}\right)^k; x = 1, 2, \dots, n \quad (4.137)$$

where  $k$  and  $n$  are constants characteristic of the underlying random phenomenon. Determine  $f(x)$ , the pdf for this random variable, and, for the specific values  $k = 2, n = 8$ , compute and plot  $f(x)$  and  $F(x)$ .

**4.7** The random variable,  $X$ , has the following pdf:

$$f(x) = \begin{cases} cx & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases} \quad (4.138)$$

- (i) First obtain the value of the constant,  $c$ , required for this to be a legitimate pdf, and then obtain an expression for the cdf  $F(x)$ .
- (ii) Obtain  $P(X \leq 1/2)$  and  $P(X \geq 1/2)$ .
- (iii) Obtain the value  $x_m$  such that

$$P(X \leq x_m) = P(X \geq x_m) \quad (4.139)$$

**4.8** From the distribution of residence times in an ideal CSTR is given in Eq (4.41), determine, for a reactor with average residence time,  $\tau = 30$  mins, the probability that a reactant molecule (i) spends *less than* 30 mins in the reactor; (ii) spends *more than* 30 mins in the reactor; (iii) spends *less than*  $(30 \ln 2)$  mins in the reactor; and (iv) spends *more than*  $(30 \ln 2)$  mins in the reactor.

### Section 4.3

**4.9** Determine  $E(X)$  for the discrete random variable in Exercise 4.5; for the continuous random variable in Exercise 4.6; and establish that  $E(X)$  for the residence time distribution in Eq (4.41) is  $\tau$ , thereby justifying why this parameter is known as the “mean residence time.”

**4.10** (Adapted from Stirzaker, 2003<sup>1</sup>) Show that  $E(X)$  exists for the discrete random variable,  $X$ , with the pdf:

$$f(x) = \frac{4}{x(x+1)(x+2)}; x = 1, 2, \dots \quad (4.140)$$

while  $E(X)$  *does not* exist for the discrete random random variable with the pdf

$$f(x) = \frac{1}{x(x+1)}; x = 1, 2, \dots \quad (4.141)$$

**4.11** Establish that  $E(X) = 1/p$  for a random variable  $X$  whose pdf is

$$f(x) = p(1-p)^{x-1}; x = 1, 2, 3, \dots \quad (4.142)$$

by differentiating with respect to  $p$  both sides of the expression:

$$\sum_{x=1}^{\infty} p(1-p)^{x-1} = 1 \quad (4.143)$$

**4.12** From the definition of the mathematical expectation function,  $E(.)$ , establish that for the random variable,  $X$ , discrete *or* continuous:

$$E[k_1g_1(X) + k_2g_2(X)] = k_1E[g_1(X)] + k_2E[g_2(X)], \quad (4.144)$$

and that given  $E(X) = \mu$ ,

$$E[(X - \mu)^3] = E(X^3) - 3\mu\sigma^2 - \mu^3 \quad (4.145)$$

---

<sup>1</sup>D. Stirzaker, (2003). *Elementary Probability*, 2<sup>nd</sup> Ed., Cambridge University Press, p120.

where  $\sigma^2$  is the variance, defined by  $\sigma^2 = \text{Var}(X) = E[(X - \mu)^2]$ .

#### Section 4.4

**4.13** Show that for two random variables  $X$  and  $Y$ , and a third random variable defined as

$$Z = X - Y \quad (4.146)$$

show, from the definition of the expectation function, that regardless of whether the random variables are continuous or discrete,

$$\begin{aligned} E(Z) &= E(X) - E(Y) \\ \text{i.e., } \mu_Z &= \mu_X - \mu_Y \end{aligned} \quad (4.147)$$

and that

$$\text{Var}(Z) = \text{Var}(X) + \text{Var}(Y) \quad (4.148)$$

when  $E[(X - \mu_X)(Y - \mu_Y)] = 0$  (i.e., when  $X$  and  $Y$  are *independent*: see Chapter 5).

**4.14** Given that the pdf of a certain discrete random variable  $X$  is:

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}; x = 0, 1, 2, \dots \quad (4.149)$$

Establish the following results:

$$\sum_{x=0}^{\infty} f(x) = 1 \quad (4.150)$$

$$E(X) = \lambda \quad (4.151)$$

$$\text{Var}(X) = \lambda \quad (4.152)$$

**4.15** Obtain the variance and skewness of the discrete random variable in Exercise 4.5 and for the continuous random variable in Exercise 4.6. Which random variable's distribution is skewed and which is symmetric?

**4.16** From the formal definitions of the moment generating function, establish Eqns (4.95) and (4.96).

**4.17** Given the pdf for the residence time for two identical CSTRs in series as

$$f(x) = \frac{1}{\tau^2} x e^{-x/\tau} \quad (4.153)$$

(i) obtain the MGF for this pdf and compare it with that derived in Example 4.7 in the text. From this comparison, what would you conjecture to be the MGF for the distribution of residence times for  $n$  identical CSTRs in series?

(ii) Obtain the characteristic function for the pdf in Eq (4.41) for the single CSTR and also for the pdf in Eq (4.153) for two CSTRs. Compare the two characteristic functions and conjecture what the corresponding characteristic function will be for the distribution of residence times for  $n$  identical CSTRs in series.

**4.18** Given that  $M(t)$  is the moment generating function of a random variable, define the “psi-function,”  $\psi(t)$ , as:

$$\psi(t) = \ln M(t) \quad (4.154)$$

- (i) Prove that  $\psi'(0) = \mu$ , and  $\psi''(0) = \sigma^2$ , where each prime  $'$  indicates differentiation with respect to  $t$ ; and  $E(X) = \mu$ , is the mean of the random variable, and  $\sigma^2$  is the variance, defined by  $\sigma^2 = \text{Var}(X) = E[(X - \mu)^2]$ .  
(ii) Given the pdf of a discrete random variable  $X$  as:

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}; x = 0, 1, 2, \dots$$

obtain its  $\psi(t)$  function and show, using the results in (i) above, that the mean and variance of this pdf are identical.

**4.19** The pdf for the yield data discussed in Chapter 1 was postulated as

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}; -\infty < y < \infty \quad (4.155)$$

If we are given that  $\mu$  is the mean, first establish that the mode is also  $\mu$ , and then use the fact that the distribution is perfectly symmetric about  $\mu$  to establish that median is also  $\mu$ , hence confirming that for this distribution, the mean, mode and median coincide.

**4.20** Given the pdf:

$$f(x) = \frac{1}{\pi} \frac{1}{1+x^2}; -\infty < x < \infty \quad (4.156)$$

find the mode and the median and show that they coincide. **For extra credit:** Establish that  $\mu = E(X)$  does not exist.

**4.21** Compute the median and the other quartiles for the random variable whose pdf is given as:

$$f(x) = \begin{cases} x & 0 < x < 2 \\ 0 & \text{otherwise} \end{cases} \quad (4.157)$$

**4.22** Given the binary random variable,  $X$ , that takes the value 1 with probability  $p$ , and the value 0 with probability  $(1 - p)$ , so that its pdf is given by

$$f(x) = \begin{cases} 1-p & x = 0; \\ p & x = 1; \\ 0 & \text{elsewhere.} \end{cases} \quad (4.158)$$

obtain an expression for the entropy  $\mathcal{H}(X)$  and show that it is maximized when  $p = 0.5$ , taking on the value  $\mathcal{H}^*(X) = 1$  at this point.

### Section 4.5

**4.23** First show that the cumulative hazard function,  $H(x)$ , for the random variable,  $X$ , the residence time in a CSTR is the linear function,

$$H(x) = \eta x \quad (4.159)$$

(where  $\eta = \frac{1}{\tau}$ ). Next, for a related random variable,  $Y$ , whose cumulative hazard function is given by

$$H(y) = (\eta y)^\zeta \quad (4.160)$$

where  $\zeta$  is a constant parameter, show that the corresponding survival function is

$$S(y) = e^{-(\eta x)^\zeta} \quad (4.161)$$

and from here obtain the pdf,  $f(y)$ , for this random variable.

**4.24** Given the pdf for the residence time for two identical CSTRs in series in Exercise 4.17, Eq (4.153), determine the survival function,  $S(x)$ , and the hazard function,  $h(x)$ . Compare them to the corresponding results obtained for the single CSTR in Example 4.8 and Example 4.9 in the text.

## APPLICATION PROBLEMS

**4.25** Before an automobile parts manufacturer takes full delivery of polymer resins made by a supplier in a reactive extrusion process, a sample is processed and the performance is tested for “Toughness.” The batch is either accepted (if the processed sample’s “Toughness” equals or exceeds  $140 \text{ J/m}^3$ ) or it is rejected. As a result of process and raw material variability, the acceptance/rejection status of each batch varies randomly. If the supplier sends four batches weekly to the parts manufacturer, and each batch is made independently on the extrusion process, so that the ultimate fate of one batch is independent of the fate of any other batch, define  $X$  as the random variable representing the number of acceptable batches a week and answer the following questions:

- (i) Obtain the sample space,  $\Omega$ , and the corresponding random variable space,  $V$ .
- (ii) First, assume equal probability of acceptance and rejection, and obtain the pdf,  $f(x)$ , for the entire sample space. If, for long term profitability it is necessary that *at least* 3 batches be acceptable per week, what is the probability that the supplier will remain profitable?

**4.26** Revisit Problem 4.25 above and consider that after an extensive process and control system improvement project, the probability of acceptance of a single batch is improved to 0.8; obtain the new pdf,  $f(x)$ . If the revenue from a single acceptable batch is \$20,000, but every rejected batch costs the supplier \$8,000 in retrieval and incineration fees, which will be deducted from the revenue, what is the expected net revenue per week under the current circumstances?

**4.27** A gas station situated on a back country road has only one gasoline pump and one attendant and, on average, receives  $\eta = 3$  (cars/hour). The average rate at which this lone attendant services the cars is  $\zeta$  (cars/hour). It can be shown that the total number of cars at this gas station at any time (i.e. the one currently being served, and those waiting in line to be served) is the random variable  $X$  with the following pdf:

$$f(x) = \left(1 - \frac{\eta}{\zeta}\right) \left(\frac{\eta}{\zeta}\right)^x ; x = 0, 1, 2, \dots \quad (4.162)$$

- (i) Show that so long as  $\eta < \zeta$ , the probability that the line at the gas station is infinitely long is *zero*.
- (ii) Find the value of  $\zeta$  required so that the expected value of the total number of

cars at the station is 2.

(iii) Using the  $\zeta$  value obtained in (ii), find the probability that there are *more than* two cars at the station, and also the probability that there are *no* cars.

**4.28** The distribution of income of families in the US in 1979 (in actual dollars uncorrected for inflation) is shown in the table below:

| Income level, $x$ ,<br>( $\times \$10^3$ ) | Percent of Population<br>with income level, $x$ |
|--|---|
| 0–5  | 4   |
| 5–10                                       | 13  |
| 10–15                                      | 17  |
| 15–20                                      | 20  |
| 20–25                                      | 16  |
| 25–30                                      | 12  |
| 30–35                                      | 7   |
| 35–40                                      | 4   |
| 40–45                                      | 3   |
| 45–50                                      | 2   |
| 50–55                                      | 1   |
| > 55                                       | 1   |

(i) Plot the data histogram and comment on the shape.

(ii) Using the center of the interval to represent each income group, determine the mean, median, mode; and the variance and skewness for this data set. Comment on how consistent the numerical values computed for these characteristics are with the shape of the histogram.

(iii) If the 1979 population is broadly classified according to income into “Lower Class” for income range (in thousands of dollars) 0–15, “Middle Class” for income range, 15–50 and “Upper Class” for income range  $> 50$ , what is the probability that two people selected at random and sequentially to participate in a survey from the Census Bureau (in preparation for the 1980 census) are (a) both from the “Lower Class,” (b) both from the “Middle Class,” (c) one from the “Middle class” and one from the “Upper class,” and (d) both from the “Upper class”?

(iv) If, in 1979, engineers with *at least* 3 years of college education (excluding graduate students) constitute approximately 1% of the population, (2.2 million out of 223 million) and span the income range from 20–55, determine the probability that an individual selected at random from the population is in the middle class given that he/she is an engineer. Determine the converse, that the person selected at random is an engineer given that he/she is in the middle class.

**4.29** Life-testing results on a first generation microprocessor-based (computer-controlled) toaster indicate that  $X$ , the life-span (in years) of the central control chip, is a random variable that is reasonably well-modeled by the pdf:

$$f(x) = \frac{1}{\beta} e^{-x/\beta}; x > 0 \quad (4.163)$$

with  $\beta = 6.25$ . A malfunctioning chip will have to be replaced to restore proper toaster operation.

(i) The warranty for the chip is to be set at  $x_w$  years (in whole integers) such that

no more than 15% would have to be replaced before the warranty period expires. Find  $x_w$ .

(ii) In planning for the second generation toaster, design engineers wish to set a target value to aim for ( $\beta = \beta_2^*$ ) such that 85% of the second generation chips survive beyond 3 years. Determine  $\beta_2^*$  and interpret your results in terms of the implied “fold increase” in mean life-span from the first to the second generation of chips.

**4.30** The probability of a single transferred embryo resulting in a live birth in an in-vitro fertilization treatment,  $p$ , is given as 0.5 for a younger patient and 0.2 for an older patient. When  $n = 5$  embryos are transferred in a single treatment, it is also known that if  $X$  is the total number of live births resulting from this treatment, then  $E(X) = 2.5$  for the younger patient and  $E(X) = 1$  for the older patient, and the associated variance,  $Var(X) = 1.25$  for the younger and  $Var(X) = 0.8$  for the older.

(i) Use Markov’s inequality and Chebyshev’s inequality to obtain bounds on the probability of each patient giving birth to quadruplets or a quintuplets at the end of the treatment.

(ii) These bounds are known to be quite conservative, but to determine just how conservative, compute the actual probabilities of the stated events for each patient given that an appropriate pdf for  $X$  is

$$f(x) = \frac{5!}{x!(5-x)!} p^x (1-p)^{5-x} \quad (4.164)$$

where  $p$  is as given above. Compare the actual probabilities with the Markov and Chebychev bounds and identify which bound is sharper.

**4.31** The following data table, obtained from the United States Life Tables 1969–71, (published in 1973 by the National Center for Health Statistics) shows the probability of survival until the age of 65 for individuals of the given age<sup>2</sup>.

| Age<br>$y$ | Prob of survival<br>to age 65 |
|------------|-------------------------------|
| 0          | 0.72                          |
| 10         | 0.74                          |
| 20         | 0.74                          |
| 30         | 0.75                          |
| 35         | 0.76                          |
| 40         | 0.77                          |
| 45         | 0.79                          |
| 50         | 0.81                          |
| 55         | 0.85                          |
| 60         | 0.90                          |

The data should be interpreted as follows: the probability that all newborns, and children up to the age of ten survive until 65 years of age is 0.72; for those older than 10 and up to 20 years, the probability of survival to 65 years is 0.74, and so on.

Assuming that the data is still valid in 1975, a community cooperative wishes to

---

<sup>2</sup>More up-to-date versions, available, for example, in *National Vital Statistics Reports*, Vol. 56, No. 9, December 28, 2007 contain far more detailed information.

set up a life insurance program that year whereby each participant pays a relatively small annual premium,  $\$a$ , and, in the event of death before 65 years, a one-time death gratuity payment of  $\$n$  is made to the participant's designated beneficiary. If the participant survives beyond 65 years, nothing is paid. If the cooperative is to realize a fixed, modest expected revenue,  $\$R_E = \$30$ , per year, per participant, over the duration of his/her participation (mostly to cover administrative and other costs) provide answers to the following questions:

- (i) For a policy based on a fixed annual premium of \$90 for all participants, and age-dependent payout, determine values for  $\pi(y)$ , the published payout for a person of age  $y$  that dies before age 65, for all values of  $y$  listed in this table.
- (ii) For a policy based instead on a fixed death payout of \$8,000, and age-dependent annual premiums, determine values for  $a(y)$ , the published annual premium to be collected each year from a participant of age  $y$ .
- (iii) If it becomes necessary to increase the expected revenue by 50% as a result of increased administrative and overhead costs, determine the effect on each of the policies in (i) and (ii) above.
- (iv) If by 1990, the probabilities of survival have increased across the board by 0.05, determine the effect on each of the policies in (i) and (ii).

# Chapter 5

## Multidimensional Random Variables

|                                |  |     |
|--------------------------------|--|-----|
| 5.1                            | Introduction and Definitions .....                                     | 137 |
| 5.1.1                          | Perspectives .....   | 138 |
| 5.1.2                          | 2-Dimensional (Bivariate) Random Variables .....                       | 139 |
| 5.1.3                          | Higher-Dimensional (Multivariate) Random Variables .....               | 140 |
| 5.2                            | Distributions of Several Random Variables .....                        | 141 |
| 5.2.1                          | Joint Distributions .....  | 141 |
| 5.2.2                          | Marginal Distributions .....   | 144 |
| 5.2.3                          | Conditional Distributions .....  | 147 |
| 5.2.4                          | General Extensions .....   | 152 |
| 5.3                            | Distributional Characteristics of Jointly Distributed Random Variables | 153 |
| 5.3.1                          | Expectations .....   | 154 |
| Marginal Expectations .....    | 155  |     |
| Conditional Expectations ..... | 156  |     |
| 5.3.2                          | Covariance and Correlation .....                                       | 157 |
| 5.3.3                          | Independence .....   | 158 |
| 5.4                            | Summary and Conclusions .....  | 163 |
|                                | REVIEW QUESTIONS .....   | 164 |
|                                | EXERCISES .....  | 166 |
|                                | APPLICATION PROBLEMS .....   | 168 |

*Servant of God, well done,  
well has thou fought the better fight,  
who single hast maintained,  
against revolted multitudes the cause of truth,  
in word mightier than they in arms.*

John Milton (1608–1674)

When the outcome of interest in an experiment is not one, but two or more variables simultaneously, additional issues arise that are not fully addressed by the probability machinery as it stands at the end of the last chapter. The concept of the random variable, restricted as it currently is to the single, one-dimensional random variable  $X$ , needs to be extended to higher dimensions; and doing so is the sole objective of this chapter. With the introduction of a few new concepts, new varieties of the probability distribution function (pdf) emerge along with new variations on familiar results; together, they expand and supplement what we already know about random variables and bring to a conclusion the discussion we started in Chapter 4.

## 5.1 Introduction and Definitions

### 5.1.1 Perspectives

Consider a clinical study of the *additional effects* of the Type 2 diabetes drug, Avandia®, in which a group of 193 patients with type 2 diabetes who had undergone cardiac bypass surgery were randomly assigned to receive the drug or a placebo. After one year, the researchers reported that patients taking Avandia® not only had better blood sugar control, they also showed improved cholesterol levels, fewer signs of inflammation of blood vessels, and lower blood pressure, compared with those on a placebo.<sup>1</sup>

Extracting the desired scientific information accurately and efficiently from the clinical study data, of course, relies on many principles of probability, statistical analysis and experimental design — issues that are not of concern at this moment. For purposes of this chapter's discussion, we restrict our attention to the basic (but central) fact that for each patient in the study, the result of interest involves not one but several variables simultaneously, including: (i) blood sugar level, (ii) cholesterol levels, (more specifically, the “low-density lipoprotein,” or LDL, version, and the “high-density lipoprotein,” or HDL, version), and (iii) blood pressure, (more specifically, the systolic and the diastolic pressures).

This is a real-life example of an experiment whose outcome is intrinsically multivariate, consisting of several distinct variables, each subject to random variability. As it currently stands, the probability machinery of Chapter 4 is only capable of dealing with one single random variable at a time. As such, we are only able to use it to characterize the variability inherent in each of the variables of interest *one at a time*. This raises some important questions that we did not have to contend with when dealing with single random variables:

1. Do these physiological variables vary *jointly* or separately? For example, do patients with high LDL cholesterol levels tend to have high systolic blood pressures also, or do the levels of one have nothing in common with the levels of the other?
2. If there is even the remotest possibility that one variable “interacts” with another, can we deal with each variable by itself as if the others do not exist without incurring serious errors?
3. If we accept that until proven otherwise these variables should be considered *jointly*, how should such joint variabilities be represented?
4. What other aspects of the joint behavior of inter-related random vari-

---

<sup>1</sup> “Avandia May Slow Atherosclerosis After Bypass Surgery”, by Steven Reinberg, *US News and World Report*, April 1, 2008.

ables provide useful means of characterizing jointly varying random variables?

These questions indicate that what we know about random variables from Chapter 4 must be extended appropriately to enable us deal with the new class of issues that arise when multiple random variables must be considered simultaneously.

The logical place to begin, of course, is with a 2-dimensional (or bivariate) random variable, before extending the discussion to the general case with  $n > 2$  variables.

### 5.1.2 2-Dimensional (Bivariate) Random Variables

The following is a direct extension of the formal definition of the single random variable given in Chapter 4.

#### **Definition: Bivariate Random Variable.**

Given a random experiment with a sample space  $\Omega$ , and a probability set function  $P(\cdot)$  defined on its subsets; let there be a function  $X$ , defined on  $\Omega$ , which assigns to each element  $\omega \in \Omega$ , one and only one ordered number pair  $(X_1(\omega), X_2(\omega))$ . This function,  $X$ , is called a two-dimensional, or bivariate *random variable*.

As with the single random variable case, associated with this two-dimensional random variable is a space,  $V$ , and a probability set function  $P_X$  induced by  $X = (X_1, X_2)$ , where  $V$  is defined as:

$$V = \{(x_1, x_2) : X_1(\omega) = x_1, X_2(\omega) = x_2; \omega \in \Omega\} \quad (5.1)$$

The most important point to note at this point is that the random variable space  $V$  involves  $X_1$  and  $X_2$  simultaneously; it is not merely a union of separate spaces  $V_1$  for  $X_1$  and  $V_2$  for  $X_2$ .

An example of a bivariate random variable was presented in Example 4.4 in Chapter 4; here is another.

#### **Example 5.1 BIVARIATE RANDOM VARIABLE AND INDUCED PROBABILITY FUNCTION FOR COIN TOSS EXPERIMENT**

Consider an experiment involving tossing a coin 2 times and recording the number of observed heads and tails: (1) Obtain the sample space  $\Omega$ ; and (2) Define  $X$  as a two-dimensional random variable  $(X_1, X_2)$  where  $X_1$  is the number of heads obtained in the first toss, and  $X_2$  is the number of heads obtained in the second toss. Obtain the new space  $V$ . (3) Assuming equiprobable outcomes, obtain the induced probability  $P_X$ .

**Solution:**

(1) From the nature of the experiment, the required sample space,  $\Omega$ , is given by

$$\Omega = \{HH, HT, TH, TT\} \quad (5.2)$$

consisting of all 4 possible outcomes, which may be represented respectively, as  $\omega_i; i = 1, 2, 3, 4$ , so that

$$\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}. \quad (5.3)$$

(2) By definition of  $X$ , we see that  $X(\omega_1) = (1, 1); X(\omega_2) = (1, 0); X(\omega_3) = (0, 1); X(\omega_4) = (0, 0)$ ; so that the space  $V$  is given by:

$$V = \{(1, 1); (1, 0); (0, 1); (0, 0)\} \quad (5.4)$$

since these are all the possible values that the two-dimensional  $X$  can take.

(3) This is a case where there is a direct one-to-one mapping between the 4 elements of the original sample space  $\Omega$  and the induced random variables space  $V$ ; as such, for equiprobable outcomes, we obtain,

$$\begin{aligned} P_X(1, 1) &= 1/4 \\ P_X(1, 0) &= 1/4 \\ P_X(0, 1) &= 1/4 \\ P_X(0, 0) &= 1/4 \end{aligned} \quad (5.5)$$

In making sense of the formal definition given here for the bivariate (2-dimensional) random variable, the reader should keep in mind the practical considerations presented in Chapter 4 for the single random variable. The same issues there apply here. In a practical sense, the bivariate random variable may be considered simply, if informally, as an experimental outcome with two components, each with numerical values that are subject to random variations with exact replicate performance of the experiment.

For example, consider a polymer used for packaging applications, for which the quality measurements of interest are “melt index” (indicative of the molecular weight distribution), and “density” (indicative of co-polymer composition). With each performance of lab analysis on samples taken from the manufacturing process, the values obtained for each of these quantities are subject to random variations. Without worrying so much about the original sample space or the induced one, we may consider the packaging polymer quality characteristics directly as the two-dimensional random variable whose components are “melt index” (as  $X_1$ ), and “density” (as  $X_2$ ).

We now note that it is fairly common for many textbooks to use  $X$  and  $Y$  to represent bivariate random variables. We choose to use  $X_1$  and  $X_2$  because it offers a notational convenience that facilitates generalization to  $n > 2$ .

### 5.1.3 Higher-Dimensional (Multivariate) Random Variables

The foregoing discussion is generalized to  $n > 2$  as follows.

**Definition: Multivariate Random Variable.**

Given a random experiment with a sample space  $\Omega$ , and a probability set function  $P(\cdot)$  defined on its subsets; let there be a function  $X$ , defined on  $\Omega$  which assigns to each element  $\omega \in \Omega$ , one and only one  $n$ -tuple  $(X_1(\omega), X_2(\omega), \dots, X_n(\omega))$  to each element  $\omega \in \Omega$ . This function,  $X$ , is called an  $n$ -dimensional *random variable*.

Similarly, associated with this  $n$ -dimensional random variable is a space,  $V$ :

$$V = \{(x_1, x_2, \dots, x_n) : X_1(\omega) = x_1, X_2(\omega) = x_2, \dots, X_n(\omega) = x_n; \omega \in \Omega\} \quad (5.6)$$

and a probability set function  $P_X$  induced by  $X$ .

As a practical matter, we may observe, for example, that in the Avandia® study mentioned at the beginning of this chapter, the outcome of interest for each patient is a continuous, 5-dimensional random variable,  $X$ , whose components are:  $X_1$  = Blood sugar level;  $X_2$  = LDL cholesterol level;  $X_3$  = HDL cholesterol level;  $X_4$  = systolic blood pressure; and  $X_5$  = diastolic blood pressure. The specific observed values for each patient will be the quintuple measurement  $(x_1, x_2, x_3, x_4, x_5)$ .

Everything we have discussed above for the bivariate random variable  $n = 2$  extends directly for the general  $n$ .

## 5.2 Distributions of Several Random Variables

### 5.2.1 Joint Distributions

The results of Example 5.1 can be written as:

$$f(x_1, x_2) = \begin{cases} 1/4; & x_1 = 1, x_2 = 1 \\ 1/4; & x_1 = 1, x_2 = 0 \\ 1/4; & x_1 = 0, x_2 = 1 \\ 1/4; & x_1 = 0, x_2 = 0 \\ 0; & \text{otherwise} \end{cases} \quad (5.7)$$

showing how the probabilities are distributed over the 2-dimensional random variable space,  $V$ . Once again, we note the following about the function  $f(x_1, x_2)$ :

- $f(x_1, x_2) > 0; \forall x_1, x_2$
- $\sum_{x_1} \sum_{x_2} f(x_1, x_2) = 1$

We may now generalize beyond this specific example as follows:

**Definition: Joint pdf**

Let there exist a sample space  $\Omega$  (along with a probability set function,  $P$ , defined on its subsets), and a random variable  $X = (X_1, X_2)$ , with an attendant random variable space  $V$ : a function  $f$  defined on  $V$  such that:

1.  $f(x_1, x_2) \geq 0; \forall x_1, x_2 \in V;$
2.  $\sum_{x_1} \sum_{x_2} f(x_1, x_2) = 1; \forall x_1, x_2 \in V;$
3.  $P_X(X_1 = x_1, X_2 = x_2) = f(x_1, x_2)$

is called the *joint* probability distribution function of the *discrete* two-dimensional random variable  $X = (X_1, X_2)$ .

These results are direct extensions of the axiomatic statements given earlier for the discrete single random variable pdf.

The probability that both  $X_1 < x_1$  and  $X_2 < x_2$  is given by the cumulative distribution function,

$$F(x_1, x_2) = P(X_1 < x_1, X_2 < x_2) \quad (5.8)$$

valid for discrete and continuous random variables. When  $F$  is a continuous function of both  $x_1$  and  $x_2$  and possesses first partial derivatives, the two-dimensional function,

$$f(x_1, x_2) = \frac{\partial}{\partial x_1} \frac{\partial}{\partial x_2} F(x_1, x_2) \quad (5.9)$$

is called the *joint* probability density function for the continuous two-dimensional random variables  $X_1$  and  $X_2$ . As with the discrete case, the formal properties of the continuous joint pdf are:

1.  $f(x_1, x_2) \geq 0; \forall x_1, x_2 \in V;$
2.  $f$  has at most a finite number of discontinuities in every finite interval in  $V$ ;
3. The double integral,  $\int_{x_1} \int_{x_2} f(x_1, x_2) dx_1 dx_2 = 1$ ;
4.  $P_X(A) = \int_A \int f(x_1, x_2) dx_1 dx_2$ ; for  $A \subset V$

Thus,

$$P(a_1 \leq X_1 \leq a_2; b_1 \leq X_2 \leq b_2) = \int_{b_1}^{b_2} \int_{a_1}^{a_2} f(x_1, x_2) dx_1 dx_2 \quad (5.10)$$

These results generalize directly to the multidimensional random variable  $X = (X_1, X_2, \dots, X_n)$  with a joint pdf  $f(x_1, x_2, \dots, x_n)$ .

**Example 5.2 JOINT PROBABILITY DISTRIBUTION OF CONTINUOUS BIVARIATE RANDOM VARIABLE**

The reliability of the temperature control system for a commercial, highly exothermic polymer reactor is known to depend on the lifetimes (in years) of the control hardware electronics,  $X_1$ , and of the control valve on the cooling water line,  $X_2$ . If one component fails, the entire control system fails. The random phenomenon in question is characterized by the two-dimensional random variable  $(X_1, X_2)$  whose joint probability distribution is given as:

$$f(x_1, x_2) = \begin{cases} \frac{1}{50} e^{-(0.2x_1+0.1x_2)}; & 0 < x_1 < \infty \\ & 0 < x_2 < \infty \\ 0 & \text{elsewhere} \end{cases} \quad (5.11)$$

- (1) Establish that this is a legitimate pdf; and (2) obtain the probability that the system lasts more than two years; (3) obtain the probability that the electronic component functions for more than 5 years and the control valve for more than 10 years.

**Solution:**

- (1) If this is a legitimate joint pdf, then the following should hold:

$$\int_0^\infty \int_0^\infty f(x_1, x_2) dx_1 dx_2 = 1 \quad (5.12)$$

In this case, we have:

$$\begin{aligned} \int_0^\infty \int_0^\infty \frac{1}{50} e^{-(0.2x_1+0.1x_2)} dx_1 dx_2 &= \frac{1}{50} \left( -5e^{-0.2x_1} \Big|_0^\infty \right) \left( -10e^{-0.1x_2} \Big|_0^\infty \right) \\ &= 1 \end{aligned} \quad (5.13)$$

We therefore conclude that the given joint pdf is legitimate.

- (2) For the system to last more than 2 years, both components must simultaneously last more than 2 year. The required probability is therefore given by:

$$P(X_1 > 2, X_2 > 2) = \int_2^\infty \int_2^\infty \frac{1}{50} e^{-(0.2x_1+0.1x_2)} dx_1 dx_2 \quad (5.14)$$

which, upon carrying out the indicated integration and simplifying, reduces to:

$$P(X_1 > 2, X_2 > 2) = e^{-0.4} e^{-0.2} = 0.67 \times 0.82 = 0.549 \quad (5.15)$$

Thus, the probability that the system lasts beyond the first two years

is 0.549.

(3) The required probability,  $P(X_1 > 5; X_2 > 10)$  is obtained as:

$$\begin{aligned} P(X_1 > 5; X_2 > 10) &= \int_{10}^{\infty} \int_5^{\infty} \frac{1}{50} e^{-(0.2x_1+0.1x_2)} dx_1 dx_2 \\ &= \left( e^{-0.2x_1} \Big|_5^{\infty} \right) \left( e^{-0.1x_2} \Big|_{10}^{\infty} \right) = (0.368)^2 \\ &= 0.135 \end{aligned} \quad (5.16)$$

The preceding discussions have established the joint pdf  $f(x_1, x_2, \dots, x_n)$  as the most direct extension of the single variable pdf  $f(x)$  of Chapter 4 to higher-dimensional random variables. However, additional distributions needed to characterize other aspects of multidimensional random variables can be derived from these joint pdfs — distributions that we had no need for in dealing with single random variables. We will discuss these new varieties of distributions first for the 2-dimensional (bivariate) random variable, and then extend the discussion to the general  $n > 2$ .

### 5.2.2 Marginal Distributions

Consider the joint pdf  $f(x_1, x_2)$  for the 2-dimensional random variable  $(X_1, X_2)$ ; it represents how probabilities are jointly distributed over the entire  $(X_1, X_2)$  plane in the random variable space. Were we to integrate over the entire range of  $X_2$  (or sum over the entire range in the discrete case), what is left is the following function of  $x_1$  in the continuous case:

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2 \quad (5.17)$$

or, in the discrete case,

$$f_1(x_1) = \sum_{x_2} f(x_1, x_2) \quad (5.18)$$

This function,  $f_1(x_1)$ , characterizes the behavior of  $X_1$  alone, by itself, regardless of what is going on with  $X_2$ .

Observe that, if one wishes to determine  $P(a_1 < X_1 < a_2)$  with  $X_2$  taking any value, by definition, this probability is determined as:

$$P(a_1 < X_1 < a_2) = \int_{a_1}^{a_2} \left( \int_{-\infty}^{\infty} f(x_1, x_2) dx_2 \right) dx_1 \quad (5.19)$$

But according to Eq (5.17), the terms in the parentheses represent  $f_1(x_1)$ , hence:

$$P(a_1 < X_1 < a_2) = \int_{a_1}^{a_2} f_1(x_1) dx_1 \quad (5.20)$$

an expression that is reminiscent of probability computations for single random variable pdfs.

The function in Eq (5.17) is known as the *marginal distribution* of  $X_1$ ; and by the same token, the marginal distribution of  $X_2$ , in the continuous case, is given by:

$$f_2(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_1, \quad (5.21)$$

obtained by “integrating out”  $X_1$  from the joint pdf of  $X_1$  and  $X_2$ ; or, in the discrete case, it is:

$$f_2(x_2) = \sum_{x_1} f(x_1, x_2) \quad (5.22)$$

These pdfs,  $f_1(x_1)$  and  $f_2(x_2)$ , respectively represent the probabilistic characteristics of *each* random variable  $X_1$  and  $X_2$  considered in isolation, as opposed to  $f(x_1, x_2)$  that represents the *joint* probabilistic characteristics when considered together. The formal definitions are given as follows:

### **Definition: Marginal pdfs**

Let  $X = (X_1, X_2)$  be a 2-dimensional random variable with a joint pdf  $f(x_1, x_2)$ ; the marginal probability distribution function of  $X_1$  alone, and of  $X_2$  alone, are defined as the following functions:

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2 \quad (5.23)$$

and

$$f_2(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_1 \quad (5.24)$$

for continuous random variables, and, for discrete random variables, as the functions:

$$f_1(x_1) = \sum_{x_2} f(x_1, x_2) \quad (5.25)$$

and

$$f_2(x_2) = \sum_{x_1} f(x_1, x_2) \quad (5.26)$$

Each marginal pdf possesses all the usual properties of pdfs, i.e., for continuous random variables,

1.  $f_1(x_1) \geq 0$ ; and  $f_2(x_2) \geq 0$
2.  $\int_{-\infty}^{\infty} f_1(x_1) dx_1 = 1$ ; and  $\int_{-\infty}^{\infty} f_2(x_2) dx_2 = 1$
3.  $P(X_1 \in A) = \int_A f_1(x_1) dx_1$ ; and  $P(X_2 \in A) = \int_A f_2(x_2) dx_2$

with the integrals are replaced with sums for the discrete case. An illustrative example follows.

### Example 5.3 MARGINAL DISTRIBUTIONS OF CONTINUOUS BIVARIATE RANDOM VARIABLE

Find the marginal distributions of the joint pdfs given in Example 5.2 for characterizing the reliability of the commercial polymer reactor's temperature control system. Recall that the component random variables are  $X_1$ , the lifetimes (in years) of the control hardware electronics, and  $X_2$ , the lifetime of the control valve on the cooling water line; the joint pdf is as given in Eq (5.11):

$$f(x_1, x_2) = \begin{cases} \frac{1}{50} e^{-(0.2x_1+0.1x_2)}; & 0 < x_1 < \infty \\ & 0 < x_2 < \infty \\ 0 & \text{elsewhere} \end{cases}$$

**Solution:**

(1) For this continuous bivariate random variable, we have from Eq (5.17) that:

$$\begin{aligned} f_1(x_1) &= \int_0^{\infty} \frac{1}{50} e^{-(0.2x_1+0.1x_2)} dx_2 \\ &= \frac{1}{50} e^{-0.2x_1} \int_0^{\infty} e^{-0.1x_2} dx_2 = \frac{1}{5} e^{-0.2x_1} \end{aligned} \quad (5.27)$$

Similarly, from Eq (5.21), we have,

$$\begin{aligned} f_2(x_2) &= \int_0^{\infty} \frac{1}{50} e^{-(0.2x_1+0.1x_2)} dx_1 \\ &= \frac{1}{50} e^{-0.1x_2} \int_0^{\infty} e^{-0.2x_1} dx_1 = \frac{1}{10} e^{-0.1x_2} \end{aligned} \quad (5.28)$$

As an exercise, the reader should confirm that each of these marginal distributions is a legitimate pdf in its own right.

These ideas extend directly to  $n > 2$  random variables whose joint pdf is given by  $f(x_1, x_2, \dots, x_n)$ . There will be  $n$  separate marginal distributions  $f_i(x_i); i = 1, 2, \dots, n$ , each obtained by integrating (or summing) out every other random variable except the one in question, i.e.,

$$f_1(x_1) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \dots, x_n) dx_2 dx_3 \cdots dx_n \quad (5.29)$$

or, in general,

$$f_i(x_i) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \dots, x_n) dx_1 dx_2 \cdots, dx_{i-1}, dx_{i+1}, \cdots dx_n \quad (5.30)$$

It is important to note that when  $n > 2$ , marginal distributions themselves can be multivariate. For example,  $f_{12}(x_1, x_2)$  is what is left of the joint pdf  $f(x_1, x_2, \dots, x_n)$  after integrating (or summing) over the remaining  $(n - 2)$  variables; it is a bivariate pdf of the two surviving random variables of interest. The concepts are simple and carry over directly; however, the notation can become quite confusing if one is not careful. We shall return to this point a bit later in this chapter.

### 5.2.3 Conditional Distributions

If the joint pdf  $f(x_1, x_2)$  of a bivariate random variable provides a description of how the two component random variables vary jointly; and if the marginal distributions  $f_1(x_1)$  and  $f_2(x_2)$  describe how each random variable behaves by itself, in isolation, without regard to the other; there remains yet one more characteristic of importance: a description of how  $X_1$  behaves *for given specific values* of  $X_2$ , and vice versa, how  $X_2$  behaves for specific values of  $X_1$  (i.e., the probability distribution of  $X_1$  *conditioned* upon  $X_2$  taking on specific values, and vice versa). Such “conditional” distributions are defined as follows:

#### Definition: Conditional pdfs

Let  $X = (X_1, X_2)$  be a 2-dimensional random variable, discrete or continuous, with a joint pdf,  $f(x_1, x_2)$ , along with marginal distributions  $f_1(x_1)$  and  $f_2(x_2)$ ; the *conditional* distribution of  $X_1$  given that  $X_2 = x_2$  is defined as:

$$f(x_1|x_2) = \frac{f(x_1, x_2)}{f_2(x_2)}; f_2(x_2) > 0 \quad (5.31)$$

Similarly, the *conditional* distribution of  $X_2$  given that  $X_1 = x_1$  is defined as:

$$f(x_2|x_1) = \frac{f(x_1, x_2)}{f_1(x_1)}; f_1(x_1) > 0 \quad (5.32)$$

The similarity between these equations and the expression for conditional probabilities of events defined as sets, as given in Eq (3.40) of Chapter 3

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (5.33)$$

should not be lost on the reader.

In Eq (5.31), the indicated pdf is a function of  $x_1$ , with  $x_2$  fixed; it is a straightforward exercise to show that this is a legitimate pdf. Observe that in the continuous case,

$$\int_{-\infty}^{\infty} f(x_1|x_2)dx_1 = \frac{\int_{-\infty}^{\infty} f(x_1, x_2)dx_1}{f_2(x_2)} \quad (5.34)$$

the numerator of which is recognized from Eq (5.21) as the marginal distribution of  $X_2$  so that:

$$\int_{-\infty}^{\infty} f(x_1|x_2)dx_1 = \frac{f_2(x_2)}{f_2(x_2)} = 1 \quad (5.35)$$

The same result holds for  $f(x_2|x_1)$  in Eq (5.32) when integrated with respect of  $x_2$ ; and, by replacing the integrals with sums, we obtain identical results for the discrete case.

#### Example 5.4 CONDITIONAL DISTRIBUTIONS OF CONTINUOUS BIVARIATE RANDOM VARIABLE

Find the conditional distributions of the 2-dimensional random variables given in Example 5.2 for the reliability of a temperature control system.

##### Solution:

Recall from the previous examples that the joint pdf is:

$$f(x_1, x_2) = \begin{cases} \frac{1}{50}e^{-(0.2x_1+0.1x_2)}; & 0 < x_1 < \infty \\ & 0 < x_2 < \infty \\ 0 & \text{elsewhere} \end{cases}$$

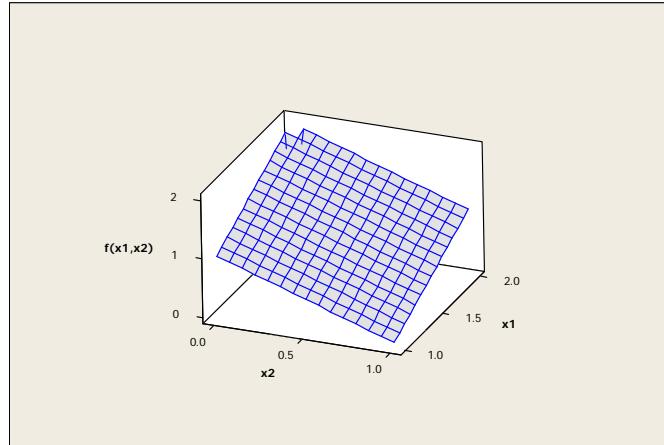
Recalling the result obtained in Example 5.3 for the marginal pdfs  $f_1(x_1)$  and  $f_2(x_2)$ , the desired conditional pdfs are given as follows:

$$\begin{aligned} f(x_1|x_2) &= \frac{\frac{1}{50}e^{-(0.2x_1+0.1x_2)}}{\frac{1}{10}e^{-0.1x_2}} \\ &= \frac{1}{5}e^{-0.2x_1} \end{aligned} \quad (5.36)$$

and for the complementary conditional pdf  $f(x_2|x_1)$ :

$$f(x_2|x_1) = \frac{\frac{1}{50}e^{-(0.2x_1+0.1x_2)}}{\frac{1}{5}e^{-0.2x_1}} = \frac{1}{10}e^{-0.1x_2} \quad (5.37)$$

The reader may have noticed two things about this specific example: (i)  $f(x_1|x_2)$  is entirely a function of  $x_1$  alone, containing no  $x_2$  whose value is to be fixed; the same is true for  $f(x_2|x_1)$  which is entirely a function of  $x_2$ , with no dependence on  $x_1$ . (ii) In fact, not only is  $f(x_1|x_2)$  a function of  $x_1$  alone; it is precisely the same function as the unconditional marginal pdf  $f_1(x_1)$  obtained earlier. The same is obtained for  $f(x_2|x_1)$ , which also turns out to



**FIGURE 5.1:** Graph of the joint pdf for the 2-dimensional random variable of Example 5.5

be the same as the unconditional marginal pdf  $f_2(x_2)$  obtained earlier. Such circumstances do not always occur for all 2-dimensional random variables, as the next example shows; but the special cases where  $f(x_1|x_2) = f_1(x_1)$  and  $f(x_2|x_1) = f_2(x_2)$  are indicative of a special relationship between the two random variables  $X_1$  and  $X_2$ , as discussed later in this chapter.

**Example 5.5 CONDITIONAL DISTRIBUTIONS OF ANOTHER CONTINUOUS BIVARIATE RANDOM VARIABLE**

Find the conditional distributions of the 2-dimensional random variables whose joint pdf is given as follows:

$$f(x_1, x_2) = \begin{cases} x_1 - x_2; & 1 < x_1 < 2 \\ 0 & 0 < x_2 < 1 \\ & \text{elsewhere} \end{cases} \quad (5.38)$$

shown graphically in Fig 5.1.

**Solution:**

To find the conditional distributions, we must first find the marginal distributions. (As an exercise, the reader may want to confirm that this joint pdf is a legitimate pdf.) These marginal distributions are obtained as follows:

$$f_1(x_1) = \int_0^1 (x_1 - x_2) dx_2 = \left( x_1 x_2 - \frac{x_2^2}{2} \right) \Big|_0^1 \quad (5.39)$$

which simplifies to give:

$$f_1(x_1) = \begin{cases} (x_1 - 0.5); & 1 < x_1 < 2 \\ 0; & \text{elsewhere} \end{cases} \quad (5.40)$$

Similarly,

$$f_2(x_2) = \int_1^2 (x_1 - x_2) dx_1 = \left( \frac{x_1^2}{2} - x_1 x_2 \right) \Big|_1^2 \quad (5.41)$$

which simplifies to give:

$$f_2(x_2) = \begin{cases} (1.5 - x_2); & 0 < x_2 < 1 \\ 0; & \text{elsewhere} \end{cases} \quad (5.42)$$

Again the reader may want to confirm that these marginal pdfs are legitimate pdfs.

With these marginal pdfs in hand, we can now determine the required conditional distributions as follows:

$$f(x_1|x_2) = \frac{(x_1 - x_2)}{(1.5 - x_2)}; \quad 1 < x_1 < 2; \quad (5.43)$$

and

$$f(x_2|x_1) = \frac{(x_1 - x_2)}{(x_1 - 0.5)}; \quad 0 < x_2 < 1; \quad (5.44)$$

(The reader should be careful to note that we did not explicitly impose the restrictive conditions  $x_2 \neq 1.5$  and  $x_1 \neq 0.5$  in the expressions given above so as to exclude the respective singularity points for  $f(x_1|x_2)$  and for  $f(x_2|x_1)$ . This is because the original space over which the joint distribution  $f(x_1, x_2)$  was defined,  $V = \{(x_1, x_2) : 1 < x_1 < 2; 0 < x_2 < 1\}$ , already excludes these otherwise troublesome points.)

Observe now that these conditional distributions show mutual dependence of  $x_1$  and  $x_2$ , unlike in Example 5.4. In particular, say for  $x_2 = 1$  (the rightmost edge of the  $x_2$ -axis of the plane in Fig 5.1), the conditional pdf  $f(x_1|x_2)$  becomes:

$$f(x_1|x_2 = 1) = 2(x_1 - 1); \quad 1 < x_1 < 2; \quad (5.45)$$

whereas, for  $x_2 = 0$  (the leftmost edge of the  $x_2$ -axis of the plane in Fig 5.1), this conditional pdf becomes

$$f(x_1|x_2 = 0) = \frac{2x_1}{3}; \quad 1 < x_1 < 2; \quad (5.46)$$

Similar arguments can be made for  $f(x_2|x_1)$  and are left as an exercise for the reader.

The following example provides a comprehensive illustration of these distributions specifically for a discrete bivariate random variable.

#### **Example 5.6 DISTRIBUTIONS OF DISCRETE BIVARIATE RANDOM VARIABLE**

An Apple® computer store in a small town stocks only three types of hardware components: “low-end”, “mid-level” and “high-end”, selling respectively for \$1600, \$2000 and \$2400; it also only stocks two types of monitors, the 20-inch type, selling for \$600, and the 23-inch type,

selling for \$900. An analysis of sales records over a 1-year period (the prices remained stable over the entire period) is shown in Table 5.1, indicating what fraction of the total sales is due to a particular hardware component and monitor type. Each recorded sale involves one hardware component *and* one monitor:  $X_1$  is the selling price of the hardware component;  $X_2$  the selling price of the accompanying monitor. The indicated “frequencies of occurrence” of each sale combination can be considered to be representative of the respective probabilities, so that Table 5.1 represents the joint distribution,  $f(x_1, x_2)$ .

**TABLE 5.1:** Joint  
pdf for computer store  
sales

| $X_2 \rightarrow$ | \$600 | \$900 |
|-------------------|-------|-------|
| $X_1 \downarrow$  |       |       |
| \$1600            | 0.30  | 0.25  |
| \$2000            | 0.20  | 0.10  |
| \$2400            | 0.10  | 0.05  |

(1) Show that  $f(x_1, x_2)$  is a legitimate pdf and find the sales combination  $(x_1^*, x_2^*)$  with the highest probability, and the one with the lowest probability.

(2) Obtain the marginal pdfs  $f_1(x_1)$  and  $f_2(x_2)$ , and from these compute  $P(X_1 = \$2000)$ , regardless of  $X_2$ , (i.e., the probability of selling a mid-level hardware component regardless of the monitor paired with it). Also obtain  $P(X_2 = \$900)$  regardless of  $X_1$ , (i.e., the probability of selling a 23-inch monitor, regardless of the hardware component with which it is paired).

(3) Obtain the conditional pdfs  $f(x_1|x_2)$  and  $f(x_2|x_1)$  and determine the highest value for each conditional probability; describe in words what each means.

**Solution:**

(1) If  $f(x_1, x_2)$  is a legitimate pdf, then it must hold that

$$\sum_{x_2} \sum_{x_1} f(x_1, x_2) = 1 \quad (5.47)$$

From the joint pdf shown in the table, this amounts to adding up all the 6 entries, a simple arithmetic exercise that yields the desired result.

The combination with the highest probability is seen to be  $X_1 = \$1600$ ;  $X_2 = \$400$  since  $P(X_1 = \$1600; X_2 = \$400) = 0.3$ ; i.e., the probability is highest (at 0.3) that any customer chosen at random would have purchased the low-end hardware (for \$1600) *and* the 20-inch monitor (for \$600). The lowest probability of 0.05 is associated with  $X_1 = \$2400$  and  $X_2 = \$900$ , i.e., the combination of a high-end hardware component and a 23-inch monitor.

(2) By definition, the marginal pdf  $f_1(x_1)$  is given by:

$$f_1(x_1) = \sum_{x_2} f(x_1, x_2) \quad (5.48)$$

so that, from the table,  $f_1(1600) = 0.3 + 0.25 = 0.55$ ; similarly,  $f_1(2000) = 0.30$  and  $f_1(2400) = 0.15$ . In the same manner, the values for  $f_2(x_2)$  are obtained as  $f_2(600) = 0.30 + 0.20 + 0.10 = 0.60$ , and  $f_2(900) = 0.4$ . These values are combined with the original joint pdf into a new Table 5.2 to provide a visual representation of the relationship between these distributions. The required probabilities are

**TABLE 5.2:** Joint and marginal pdfs for computer store sales

| $X_2 \rightarrow$ |            | \$600 | \$900 | $f_1(x_1)$ |
|-------------------|------------|-------|-------|------------|
| $X_1 \downarrow$  |            | 0.30  | 0.25  |            |
| \$1600            |            | 0.30  | 0.25  | 0.55       |
| \$2000            |            | 0.20  | 0.10  | 0.30       |
| \$2400            |            | 0.10  | 0.05  | 0.15       |
|                   | $f_2(x_2)$ | 0.6   | 0.4   | (1.0)      |

obtained directly from this table as follows:

$$P(X_1 = \$2000) = f_1(2000) = 0.30 \quad (5.49)$$

$$P(X_2 = \$900) = f_2(900) = 0.40 \quad (5.50)$$

(3) By definition, the desired conditional pdfs are given as follows:

$$f(x_1|x_2) = \frac{f(x_1, x_2)}{f_2(x_2)}; \text{ and } f(x_2|x_1) = \frac{f(x_1, x_2)}{f_1(x_1)} \quad (5.51)$$

and upon carrying out the indicated divisions using the numbers contained in Table 5.2, we obtain the result shown in Table 5.3 for  $f(x_1|x_2)$ , and in Table 5.4 for  $f(x_2|x_1)$ . From these tables, we obtain the highest conditional probability for  $f(x_1|x_2)$  as 0.625, corresponding to the probability of a customer buying the low end hardware component ( $X_1 = \$1600$ ) conditioned upon having bought the 23-inch monitor ( $X_2 = \$900$ ); i.e., in the entire population of those who bought the 23-inch monitor, the probability is highest at 0.625 that a low-end hardware component was purchased to go along with the monitor. When the conditioning variable is the hardware component, the highest conditional probability  $f(x_2|x_1)$  is a tie at 0.667 for customers buying the 20-inch monitor ( $X_2 = \$600$ ) conditioned upon buying the mid-range hardware component ( $X_1 = \$2000$ ), and those buying the high-end hardware component ( $X_1 = \$2400$ ).

**TABLE 5.3:** Conditional pdf  $f(x_1|x_2)$  for computer store sales

| $X_1$     | $f(x_1 x_2 = 600)$ | $f(x_1 x_2 = 900)$ |
|-----------|--------------------|--------------------|
| \$1600    | 0.500              | 0.625              |
| \$2000    | 0.333              | 0.250              |
| \$2400    | 0.167              | 0.125              |
| Sum Total | 1.000              | 1.000              |

**TABLE 5.4:** Conditional pdf  $f(x_2|x_1)$  for computer store sales

| $X_2 \rightarrow$   | \$600 | \$900 | Sum Total |
|---------------------|-------|-------|-----------|
| $f(x_2 x_1 = 1600)$ | 0.545 | 0.455 | 1.000     |
| $f(x_2 x_1 = 2000)$ | 0.667 | 0.333 | 1.000     |
| $f(x_2 x_1 = 2400)$ | 0.667 | 0.333 | 1.000     |

### 5.2.4 General Extensions

As noted in the section on marginal distributions, it is conceptually straightforward to extend the foregoing ideas and results to the general case with  $n > 2$ . Such a general discussion, however, is susceptible to confusion primarily because the notation can become muddled very quickly. Observe that not only can the variables whose conditional distributions we seek be multivariate, the conditioning variables themselves can also be multivariate (so that the required marginal distributions are multivariate pdfs); and there is always the possibility that there will be some variables left over that are neither of interest, nor in the conditioning set.

To illustrate, consider the 5-dimensional random variable associated with the Avandia clinical test: the primary point of concern that precipitated this study was not so much the effectiveness of the drug in controlling blood sugar; it is the potential adverse side-effect on cardiovascular function. Thus, the researchers may well be concerned with characterizing the pdf for blood pressure  $X_4, X_5$ , conditioned upon cholesterol level  $X_2, X_3$ , leaving out  $X_1$ , blood sugar level. Note how the variable of interest is bivariate, as is the conditioning variable. In this case, the desired conditional pdf is obtained as:

$$f(x_4, x_5|x_2, x_3) = \frac{f(x_1, x_2, x_3, x_4, x_5)}{f_{23}(x_2, x_3)} \quad (5.52)$$

where  $f_{23}(x_2, x_3)$  is the bivariate joint marginal pdf for cholesterol level. We see therefore that the principles transfer quite directly, and, when dealing with specific cases in practice (as we have just done), there is usually no confusion. The challenge is how to generalize without confusion.

To present the results in a general fashion and avoid confusion requires adopting a different notation: using the vector  $\mathbf{X}$  to represent the entire collection of random variables, i.e.,  $\mathbf{X} = (X_1, X_2, \dots, X_n)$ , and then partitioning this into three distinct vectors:  $\mathbf{X}^*$ , the variables of interest ( $X_4, X_5$  in the Avandia example given above);  $\mathbf{Y}$ , the conditioning variables ( $X_2, X_3$  in the Avandia example), and  $\mathbf{Z}$ , the remaining variables, if any. With this notation, we now have

$$f(\mathbf{x}^*|\mathbf{y}) = \frac{f(\mathbf{x}^*, \mathbf{y}, \mathbf{z})}{f_{\mathbf{y}}(\mathbf{y})} \quad (5.53)$$

as the most general multivariate conditional distribution.

### 5.3 Distributional Characteristics of Jointly Distributed Random Variables

The concepts of mathematical expectation and moments used to characterize the distribution of single random variables in Chapter 4 can be extended to multivariate, jointly distributed random variables. Even though we now have many more versions of pdfs to consider (joint, marginal and conditional), the primary notions remain the same.

#### 5.3.1 Expectations

The mathematical expectation of the function  $U(\mathbf{X}) = U(X_1, X_2, \dots, X_n)$  of an  $n$ -dimensional continuous random variable with joint pdf  $f(x_1, x_2, \dots, x_n)$  is given by:

$$E[U(X)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} U(x_1, x_2, \dots, x_n) f(x_1, x_2, \dots, x_n) dx_1 dx_2 \cdots dx_n \quad (5.54)$$

a direct extension of the single variable definition. The discrete counterpart is:

$$E[U(X)] = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_n} U(x_1, x_2, \dots, x_n) f(x_1, x_2, \dots, x_n) \quad (5.55)$$

#### Example 5.7 EXPECTATIONS OF CONTINUOUS BIVARIATE RANDOM VARIABLE

From the joint pdf given in Example 5.2 for the reliability of the reactor temperature control system, deduce which component is *expected* to fail first and by how long it is expected to be outlasted by the more durable component.

##### Solution:

Recalling that the random variables for this system are  $X_1$ , the lifetime (in years) of the control hardware electronics, and  $X_2$ , the lifetime of the control valve on the cooling water line, observe that the function  $U(X_1, X_2)$ , defined as

$$U(X_1, X_2) = X_1 - X_2 \quad (5.56)$$

represents the *differential* lifetimes of the two components; its expected value provides the answer to both aspects of this question as follows: By the definition of expectations,

$$E[U(X_1 - X_2)] = \frac{1}{50} \int_0^{\infty} \int_0^{\infty} (x_1 - x_2) e^{-(0.2x_1 + 0.1x_2)} dx_1 dx_2 \quad (5.57)$$

The indicated integrals can be evaluated several different ways. By expanding this expression into the difference of two double integrals as

suggested by the multiplying  $(x_1 - x_2)$ , integrating out  $x_2$  in the first and  $x_1$  in the second, leads to:

$$E[U(X_1 - X_2)] = \left( \frac{1}{5} \int_0^\infty x_1 e^{-0.2x_1} dx_1 \right) - \left( \frac{1}{10} \int_0^\infty x_2 e^{-0.1x_2} dx_2 \right); \quad (5.58)$$

and upon carrying out the indicated integration by parts, we obtain:

$$E(X_1 - X_2) = 5 - 10 = -5. \quad (5.59)$$

The immediate implication is that the expected lifetime differential favors the control valve (lifetime  $X_2$ ) so that the control hardware electronic component is expected to fail first, with the control valve expected to outlast it by 5 years.

#### Example 5.8 EXPECTATIONS OF DISCRETE BIVARIATE RANDOM VARIABLE

From the joint pdf given in Example 5.6 for the Apple computer store sales, obtain the expected revenue from each recorded sale.

##### Solution:

Recall that for this problem, the random variables of interest are  $X_1$ , the cost of the computer hardware component, and  $X_2$ , the cost of the monitor in each recorded sale. The appropriate function  $U(X_1, X_2)$ , in this case is

$$U(X_1, X_2) = X_1 + X_2 \quad (5.60)$$

the total amount of money realized on each sale. By the definition of expectations for the discrete bivariate random variable, we have

$$E[U(X_1, X_2)] = \sum_{x_2} \sum_{x_1} (x_1 + x_2) f(x_1, x_2) \quad (5.61)$$

From Table 5.1, this is obtained as:

$$\begin{aligned} E(X_1 + X_2) &= \left( \sum_{x_1} x_1 f(x_1, x_2) \right) + \left( \sum_{x_2} x_2 f(x_1, x_2) \right) \\ &= (0.55 \times 1600 + 0.30 \times 2000 + 0.15 \times 2400) \\ &\quad + (0.60 \times 600 + 0.40 \times 900) = 2560 \end{aligned} \quad (5.62)$$

so that the required expected revenue from each sale is \$2560.

In the special case where  $U(\mathbf{X}) = e^{(t_1 X_1 + t_2 X_2)}$ , the expectation,  $E[U(\mathbf{X})]$  is the joint moment generating function,  $M(t_1, t_2)$ , for the bivariate random variable  $\mathbf{X} = (X_1, X_2)$  defined by

$$M(t_1, t_2) = E[e^{(t_1 X_1 + t_2 X_2)}] = \begin{cases} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{(t_1 X_1 + t_2 X_2)} f(x_1, x_2) dx_1 dx_2; \\ \sum_{x_1} \sum_{x_2} e^{(t_1 X_1 + t_2 X_2)} f(x_1, x_2); \end{cases} \quad (5.63)$$

for the continuous and the discrete cases, respectively — an expression that generalizes directly for the  $n$ -dimensional random variable.

### Marginal Expectations

Recall that for the general  $n$ -dimensional random variable  $\mathbf{X} = (X_1, X_2, \dots, X_n)$ , the single variable marginal distribution  $f_i(x_i)$  is the distribution of the component random variable  $X_i$  alone, as if the others did not exist. It is therefore similar to the single random variable pdf dealt with extensively in Chapter 4. As such, the marginal expectation of  $U(X_i)$  is precisely as defined in Chapter 4, i.e.,

$$E[U(X_i)] = \int_{-\infty}^{\infty} U(x_i) f_i(x_i) dx_i \quad (5.64)$$

for the continuous case, and, for the discrete case,

$$E[U(X_i)] = \sum_{x_i} U(x_i) f_i(x_i) \quad (5.65)$$

In particular, when  $U(X_i) = X_i$ , we obtain the marginal mean  $\mu_{X_i}$ , i.e.,

$$E(X_i) = \mu_{X_i} = \begin{cases} \int_{-\infty}^{\infty} x_i f_i(x_i) dx_i; & \text{continuous } X_i \\ \sum_{x_i} x_i f_i(x_i); & \text{discrete } X_i \end{cases} \quad (5.66)$$

All the moments (central and ordinary) defined for the single random variable are precisely the same as the corresponding marginal moments for the multi-dimensional random variable. In particular, the marginal variance is defined as

$$\sigma_{X_i}^2 = E[(X_i - \mu_{X_i})^2] = \begin{cases} \int_{-\infty}^{\infty} (x_i - \mu_{X_i})^2 f_i(x_i) dx_i; & \text{continuous } X_i \\ \sum_{x_i} (x_i - \mu_{X_i})^2 f_i(x_i); & \text{discrete } X_i \end{cases} \quad (5.67)$$

From the expression given for the joint MGF above in Eq (5.63), observe that:

$$M(t_1, 0) = E[e^{t_1 X_1}] \quad (5.68)$$

$$M(0, t_2) = E[e^{t_2 X_2}] \quad (5.69)$$

are, respectively, the marginal MGFs for  $f_1(x_1)$  and for  $f_2(x_2)$ .

Keep in mind that in the general case, marginal distributions can be multivariate; in this case, the context of the problem at hand will make clear what such a joint-marginal distribution will look like after the remaining variables have been integrated out.

### Conditional Expectations

As in the discussion about conditional distributions, it is best to deal with the bivariate conditional expectations first. For the bivariate random variable

$\mathbf{X} = (X_1, X_2)$ , the conditional expectation  $E[U(X_1)|X_2]$  (i.e the expectation of the function  $U(X_1)$  conditioned upon  $X_2 = x_2$ ) is obtained from the conditional distribution as follows:

$$E[U(X_1)|X_2] = \begin{cases} \int_{-\infty}^{\infty} U(x_1)f(x_1|x_2)dx_1; & \text{continuous } \mathbf{X} \\ \sum_{x_1} U(x_1)f(x_1|x_2); & \text{discrete } \mathbf{X} \end{cases} \quad (5.70)$$

with a corresponding expression for  $E[U(X_2)|X_1]$  based on the conditional distribution  $f(x_2|x_1)$ . In particular, when  $U(X_1) = X_1$  (or,  $U(X_2) = X_2$ ), the result is the *conditional mean* defined by:

$$E(X_1|X_2) = \mu_{X_1|x_2} = \begin{cases} \int_{-\infty}^{\infty} x_1 f(x_1|x_2)dx_1; & \text{continuous } \mathbf{X} \\ \sum_{x_i} x_i f(x_1|x_2); & \text{discrete } \mathbf{X} \end{cases} \quad (5.71)$$

with a matching corresponding expression for  $\mu_{X_2|x_1}$ .

Similarly, if

$$U(X_1) = (X_1 - \mu_{X_1|x_2})^2 \quad (5.72)$$

we obtain the conditional variance,  $\sigma_{X_1|x_2}^2$  as:

$$\sigma_{X_1|x_2}^2 = E[(X_1 - \mu_{X_1|x_2})^2] = \begin{cases} \int_{-\infty}^{\infty} (x_1 - \mu_{X_1|x_2})^2 f(x_1|x_2)dx_1; & \text{continuous } \mathbf{X} \\ \sum_{x_1} (x_1 - \mu_{X_1|x_2})^2 f(x_1|x_2); & \text{discrete } \mathbf{X} \end{cases} \quad (5.73)$$

respectively for the continuous and discrete cases.

These concepts can be extended quite directly to general  $n$ -dimensional random variables; but, as noted earlier, one must be careful to avoid confusing notation.

### 5.3.2 Covariance and Correlation

Consider the 2-dimensional random variable  $\mathbf{X} = (X_1, X_2)$  whose marginal means are given by  $\mu_{X_1}$  and  $\mu_{X_2}$ , and respective marginal variances  $\sigma_1^2$  and  $\sigma_2^2$ ; the quantity

$$\sigma_{12} = E[(X_1 - \mu_{X_1})(X_2 - \mu_{X_2})] \quad (5.74)$$

is known as the covariance of  $X_1$  with respect to  $X_2$ ; it is a measure of the mutual dependence of variations in  $X_1$  and in  $X_2$ . It is straightforward to show from Eq (5.74) that

$$\sigma_{12} = E(X_1 X_2) - \mu_{X_1} \mu_{X_2} \quad (5.75)$$

A popular and more frequently used measure of this mutual dependence is the scaled quantity:

$$\rho = \frac{\sigma_{12}}{\sigma_1 \sigma_2} \quad (5.76)$$

where  $\sigma_1$  and  $\sigma_2$  are the positive square roots of the respective marginal variances of  $X_1$  and  $X_2$ .  $\rho$  is known as the correlation coefficient, with the attractive property that

$$-1 \leq \rho \leq 1 \quad (5.77)$$

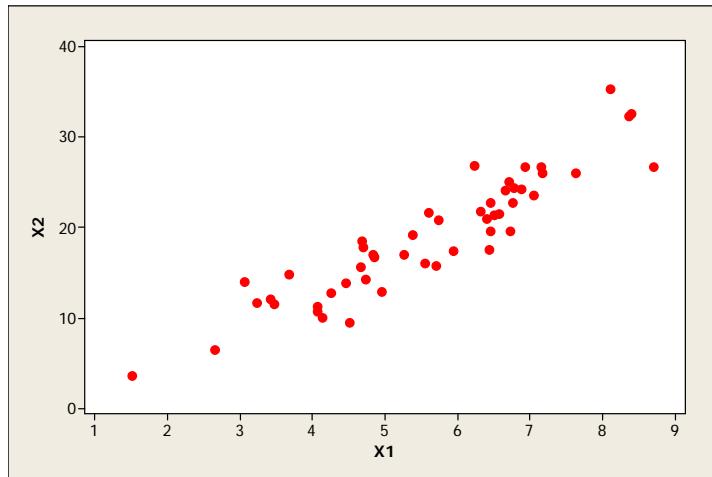
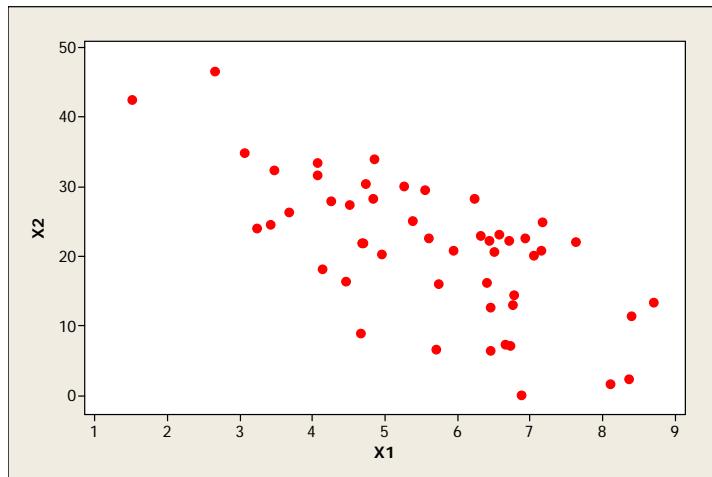
The most important points to note about the covariance,  $\sigma_{12}$ , or the correlation coefficient, are as follows:

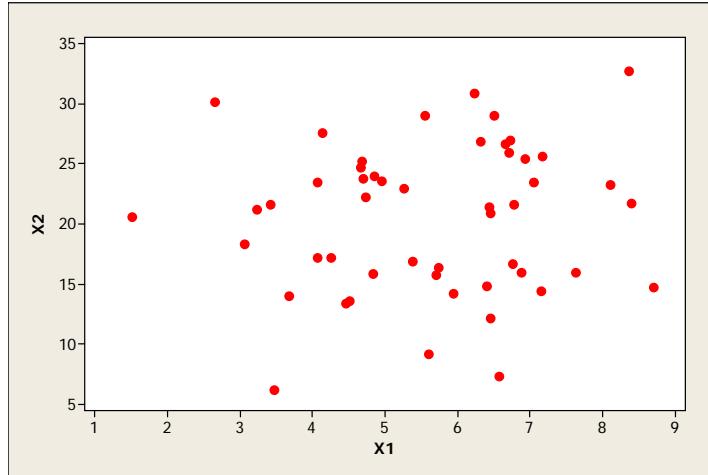
1.  $\sigma_{12}$  will be positive if values of  $X_1 > \mu_{X_1}$  are *generally* associated with values of  $X_2 > \mu_{X_2}$ , or when values of  $X_1 < \mu_{X_1}$  tend to be associated with values of  $X_2 < \mu_{X_2}$ . Such variables are said to be *positively* correlated and  $\rho$  will be positive ( $\rho > 0$ ), with the strength of the correlation indicated by the absolute value of  $\rho$ : weakly correlated variables will have low values close to zero while strongly correlated variables will have values close to 1. (See Fig 5.2.) For perfectly positively correlated variables,  $\rho = 1$ .
2. The reverse is the case when  $\sigma_{12}$  is negative: for such variables, values of  $X_1 > \mu_{X_1}$  appear preferentially together with values of  $X_2 < \mu_{X_2}$ , or else values of  $X_1 < \mu_{X_1}$  tend to be associated more with values of  $X_2 > \mu_{X_2}$ . In this case, the variables are said to be *negatively* correlated and  $\rho$  will be negative ( $\rho < 0$ ); once again, with the strength of correlation indicated by the absolute values of  $\rho$ . (See Fig 5.3). For perfectly negatively correlated variables,  $\rho = -1$ .
3. If the behavior of  $X_1$  has little or no bearing with that of  $X_2$ , as one might expect,  $\sigma_{12}$  and  $\rho$  will tend to be close to zero (See Fig 5.4); and when the two random variables are completely “independent” of each other, then both  $\sigma_{12}$  and  $\rho$  will be exactly zero.

This last point brings up the concept of stochastic independence.

### 5.3.3 Independence

Consider a situation where electronic component parts manufactured at two different plant sites are labeled “1” for plant site 1, and “2” for the other. After combining these parts into one lot, each part is drawn at random and tested: if found defective, it is labeled “0”; otherwise it is labeled “1”. Now consider the 2-dimensional random variable  $\mathbf{X} = (X_1, X_2)$  where  $X_1$  is the location of the manufacturing site (1 or 2), and  $X_2$  is the “after-test” status of the electronic component part (0 or 1). If after many such draws and tests, we discover that whether or not the part is defective has absolutely nothing to do with where it was manufactured, (i.e., a defective part is just as likely to come from one plant site as the other), we say that  $X_1$  is *independent* of  $X_2$ . A formal definition now follows:

FIGURE 5.2: Positively correlated variables:  $\rho = 0.923$ FIGURE 5.3: Negatively correlated variables:  $\rho = -0.689$

FIGURE 5.4: Essentially uncorrelated variables:  $\rho = 0.085$ **Definition: Stochastic Independence**

Let  $X = (X_1, X_2)$  be a 2-dimensional random variable, discrete or continuous;  $X_1$  and  $X_2$  are independent if the following conditions hold:

1.  $f(x_2|x_1) = f_2(x_2);$
2.  $f(x_1|x_2) = f_1(x_1);$  and
3.  $f(x_1, x_2) = f_1(x_1)f_2(x_2)$

The first point indicates that the distribution of  $X_2$  conditional on  $X_1$  is identical to the unconditional (or marginal) distribution of  $X_2$ . In other words, conditioning on  $X_1$  has no effect on the distribution of  $X_2$ , indicating that  $X_2$  is *independent of  $X_1$* . However, this very fact (that  $X_2$  is independent of  $X_1$ ) also *immediately implies* the converse: that  $X_1$  is independent of  $X_2$  (i.e., that the independence in this case is mutual). To establish this, we note that, by definition, in Eq (5.32),

$$f(x_2|x_1) = \frac{f(x_1, x_2)}{f_1(x_1)} \quad (5.78)$$

However, when  $X_2$  is independent of  $X_1$ ,

$$f(x_2|x_1) = f_2(x_2) \quad (5.79)$$

i.e., point 1 above, holds; as a consequence, by replacing  $f(x_2|x_1)$  in Eq (5.78)

above with  $f_2(x_2)$ , we obtain:

$$f(x_1, x_2) = f_1(x_1)f_2(x_2) \quad (5.80)$$

which, first of all, is item 3 in the definition above, but just as importantly, when substituted into the numerator of the expression in Eq (5.31), i.e.,

$$f(x_1|x_2) = \frac{f(x_1, x_2)}{f_2(x_2)}$$

when the conditioning is now on  $X_2$ , reduces this equation to

$$f(x_1|x_2) = f_1(x_1) \quad (5.81)$$

which is item number 2 above — indicating that  $X_1$  is *also* independent of  $X_2$ . The two variables are therefore said to be “mutually stochastically independent.”

Let us now return to a point made earlier after Example 5.4. There we noted that the distributional characteristics of the random variables  $X_1$ , the lifetime (in years) of the control hardware electronics, and  $X_2$ , the lifetime of the control valve on the cooling water line, were such that they satisfied conditions now recognizable as the ones given in points 1 and 2 above. It is therefore now clear that the special relationship between these two random variables alluded to back then is that they are stochastically independent. Note that the joint pdf,  $f(x_1, x_2)$ , for this system is a product of the two marginal pdfs, as in condition 3 above. This is *not* the case for the random variables in Example 5.5.

The following example takes us back to yet another example encountered earlier.

#### Example 5.9 INDEPENDENCE OF TWO DISCRETE RANDOM VARIABLES

Return to the two-coin toss experiment discussed in Example 5.1. From the joint pdf obtained for this bivariate random variable (given in Eq (5.7)), show that the two random variables,  $X_1$  (the number of heads obtained in the first toss), and  $X_2$  (the number of heads obtained in the second toss), are independent.

##### Solution:

By definition, and from the results in that example, the marginal distributions are obtained as follows:

$$\begin{aligned} f_1(x_1) &= \sum_{x_2} f(x_1, x_2) \\ &= f(x_1, 0) + f(x_1, 1) \end{aligned} \quad (5.82)$$

so that  $f_1(0) = 1/2$ ;  $f_1(1) = 1/2$ . Similarly,

$$\begin{aligned} f_2(x_2) &= \sum_{x_1} f(x_1, x_2) \\ &= f(0, x_2) + f(1, x_2) = 1/2 \end{aligned} \quad (5.83)$$

**TABLE 5.5:** Joint and marginal pdfs for two-coin toss problem of Example 5.1

|                  |  | $X_2 \rightarrow$ | 0   | 1   | $f_1(x_1)$ |
|------------------|--|-------------------|-----|-----|------------|
| $X_1 \downarrow$ |  | 0                 | 1/4 | 1/4 | 1/2        |
|                  |  | 1                 | 1/4 | 1/4 | 1/2        |
|                  |  | $f_2(x_2)$        | 1/2 | 1/2 | 1          |

so that  $f_2(0) = 1/2$ ;  $f_2(1) = 1/2$ ; i.e.,

$$f_1(x_1) = \begin{cases} 1/2; & x_1 = 0 \\ 1/2; & x_1 = 1 \\ 0; & \text{otherwise} \end{cases} \quad (5.84)$$

$$f_2(x_2) = \begin{cases} 1/2; & x_2 = 0 \\ 1/2; & x_2 = 1 \\ 0; & \text{otherwise} \end{cases} \quad (5.85)$$

If we now tabulate the joint pdf and the marginal pdfs, we obtain the result in Table 5.5. It is now clear that for all  $x_1$  and  $x_2$ ,

$$f(x_1, x_2) = f_1(x_1)f_2(x_2) \quad (5.86)$$

so that these two random variables are independent.

Of course, we know intuitively that the number of heads obtained in the first toss should have no effect on the number of heads obtained in the second toss, but this fact has now been established theoretically.

The concept of independence is central to a great deal of the strategies for solving problems involving random phenomena. The ideas presented in this section are therefore used repeatedly in upcoming chapters in developing models, and in solving many practical problems.

The following is one additional consequence of stochastic independence. If  $X_1$  and  $X_2$  are independent, then

$$E[U(X_1)G(X_2)] = E[U(X_1)]E[G(X_2)] \quad (5.87)$$

An immediate consequence of this fact is that in this case,

$$\sigma_{12} = 0 \quad (5.88)$$

$$\rho = 0 \quad (5.89)$$

since, by definition,

$$\sigma_{12} = E[(X_1 - \mu_{X_1})(X_2 - \mu_{X_2})] \quad (5.90)$$

and, by virtue of Eq (5.87), independence implies:

$$\sigma_{12} = E[(X_1 - \mu_{X_1})].E[(X_2 - \mu_{X_2})] = 0 \quad (5.91)$$

It also follows that  $\rho = 0$  since it is  $\sigma_{12}/\sigma_1\sigma_2$ .

A note of caution: it is possible for  $E[U(X_1)G(X_2)]$  to equal the product of expectations,  $E[U(X_1)]E[G(X_2)]$  by chance, without  $X_1$  and  $X_2$  being independent; however, if  $X_1$  and  $X_2$  are independent, then Eq. (5.87) will hold. This expression is therefore a necessary but not sufficient condition.

We must exercise care in extending the definition of stochastic independence to the  $n$ -dimensional random variable where  $n > 2$ . The random variables  $X_1, X_2, \dots, X_n$  are said to be mutually stochastically independent if and only if,

$$f(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f_i(x_i) \quad (5.92)$$

where  $f(x_1, x_2, \dots, x_n)$  is the joint pdf, and  $f_i(x_i); i = 1, 2, \dots, n$ , are the  $n$  individual marginal pdfs. On the other hand, these random variables are *pairwise* stochastically independent if every pair  $X_i, X_j; i \neq j$ , is stochastically independent.

Obviously, mutual stochastic independence implies pairwise stochastic independence, but not vice versa.

## 5.4 Summary and Conclusions

The primary objective of this chapter was to extend the ideas presented in Chapter 4 for the single random variable to the multidimensional case, where the outcome of interest involves two or more random variables simultaneously. With such higher-dimensional random variables, it became necessary to introduce a new variety of pdfs different from, but still related to, the familiar one encountered in Chapter 4: the *joint* pdf to characterize joint variation among the variables; the *marginal* pdfs to characterize individual behavior of each variable in isolation from others; and the *conditional* pdfs, to characterize the behavior of one random variable conditioned upon fixing the others at pre-specified values. This new array of pdfs provide the full set of mathematical tools for characterizing various aspects of multivariate random variables much as the  $f(x)$  of Chapter 4 did for single random variables.

The possibility of two or more random variables “co-varying” simultaneously, which was not of concern with single random variables, led to the introduction of two additional and related quantities, co-variance and correlation, with which one quantifies the mutual dependence of two random variables. This in turn led to the important concept of stochastic independence, that one random variable is entirely unaffected by another. As we shall see in subsequent chapters, when dealing with multiple random variables, the analysis of joint behavior is considerably simplified if the random variables in question

are independent. We shall therefore have cause to recall some of the results of this chapter at that time.

Here are some of the main points of this chapter again.

- A multivariate random variable is defined in the same manner as a single random variable, but the associated space,  $V$ , is higher-dimensional;
- The joint pdf of a bivariate random variable,  $f(x_1, x_2)$ , shows how the probabilities are distributed over the two-dimensional random variable space; the joint cdf,  $F(x_1, x_2)$ , represents the probability,  $P(X_1 < x_1; X_2 < x_2)$ ; they both extend directly to higher-dimensional random variables.
- In addition to the joint pdf, two other pdfs are needed to characterize multi-dimensional random variables fully:
  - *Marginal pdf*:  $f_i(x_i)$  characterizes the individual behavior of each random variable,  $X_i$ , by itself, regardless of the others;
  - *Conditional pdf*:  $f(x_i|x_j)$  characterizes the behavior of  $X_i$  conditioned upon  $X_j$  taking on specific values.

These pdfs can be used to obtain such random variable characteristics as joint, marginal and conditional expectations.

- The covariance of two random variables,  $X_1$  and  $X_2$ , defined as

$$\sigma_{12} = E[(X_1 - \mu_{X_1})(X_2 - \mu_{X_2})]$$

(where  $\mu_{X_1}$  and  $\mu_{X_2}$ , are respective marginal expectations), provides a measure of the mutual dependence of variations in  $X_1$  and  $X_2$ . The related correlation coefficient, the scaled quantity:

$$\rho = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$$

(where  $\sigma_1$  and  $\sigma_2$  are the positive square roots of the respective marginal variances of  $X_1$  and  $X_2$ ), has the property that  $-1 \leq \rho \leq 1$ , with  $|\rho|$  indicating the strength of the mutual dependence, and the sign indicating the direction (negative or positive).

- Two random variables,  $X_1$  and  $X_2$ , are independent if the behavior of one has no bearing on the behavior of the other; more formally,

$$f(x_1|x_2) = f_1(x_1); \quad f(x_2|x_1) = f_2(x_2);$$

so that,

$$f(x_1, x_2) = f(x_1)f(x_2)$$

## REVIEW QUESTIONS

1. What characteristic of the Avandia® clinical test makes it relevant to the discussion of this chapter?
2. How many random variables at a time can the probability machinery of Chapter 4 deal with?
3. In dealing with several random variables simultaneously, what are some of the questions to be considered that were not of concern when dealing with single random variables in Chapter 4?
4. Define a bivariate random variable formally.
5. Informally, what is a bivariate random variable?
6. Define a multivariate random variable formally.
7. State the axiomatic definition of the joint pdf of a discrete bivariate random variable and of its continuous counterpart.
8. What is the general relationship between the cdf,  $F(x_1, x_2)$ , of a continuous bivariate random variable and its pdf,  $f(x_1, x_2)$ ? What conditions must be satisfied for this relationship to exist?
9. Define the marginal distributions,  $f_1(x_1)$  and  $f_2(x_2)$ , for a two-dimensional random variable with a joint pdf  $f(x_1, x_2)$ .
10. Do marginal pdfs possess the usual properties of pdfs or are they different?
11. Given a bivariate joint pdf,  $f(x_1, x_2)$ , define the conditional pdfs,  $f(x_1|x_2)$  and  $f(x_2|x_1)$ .
12. In what way is the definition of a conditional pdf similar to the conditional probability of events  $A$  and  $B$  defined on a sample space,  $\Omega$ ?
13. Define the expectation,  $E[(U(X_1, X_2))]$ , for a bivariate random variable. Extend this to an  $n$ -dimensional (multivariate) random variable.
14. Define the marginal expectation,  $E[(U(X_i))]$ , for a bivariate random variable. Extend this to an  $n$ -dimensional (multivariate) random variable.
15. Define the conditional expectations,  $E[(U(X_1)|X_2)]$  and  $E[(U(X_2)|X_1)]$ , for a bivariate random variable.
16. Given two random variables,  $X_1$  and  $X_2$ , define their covariance.
17. What is the relationship between covariance and the correlation coefficient?

- 18.** What does a negative correlation coefficient indicate about the relationship between two random variables,  $X_1$  and  $X_2$ ? What does a positive correlation coefficient indicate?
- 19.** If the behavior of the random variable,  $X_1$ , has little bearing on that of  $X_2$ , how will this manifest in the value of the correlation coefficient,  $\rho$ ?
- 20.** When the correlation coefficient of two random variables,  $X_1$  and  $X_2$ , is such that  $|\rho| \approx 1$ , what does this indicate about the random variables?
- 21.** What does it mean that two random variables,  $X_1$  and  $X_2$ , are stochastically independent?
- 22.** If two random variables are independent, what is the value of their covariance, and of their correlation coefficient?
- 23.** When dealing with  $n > 2$  random variables, what is the difference between pairwise stochastic independence and mutual stochastic independence? Does one always imply the other?

## EXERCISES

### Sections 5.1 and 5.2

**5.1** Revisit Example 5.1 in the text and define the two-dimensional random variable  $(X_1, X_2)$  as follows:  $X_1$  is the total number of heads, and  $X_2$  is the total number of tails. Obtain the space,  $V$ , and determine the complete pdf,  $f(x_1, x_2)$ , for  $x_1 = 0, 1, 2; x_2 = 0, 1, 2$ , assuming equiprobable outcomes in the original sample space.

**5.2** The two-dimensional random variable  $(X_1, X_2)$  has the following joint pdf:

$$\begin{aligned} f(1, 1) &= \frac{1}{4}; & f(2, 1) &= \frac{3}{8} \\ f(1, 2) &= \frac{1}{8}; & f(2, 2) &= \frac{1}{8} \\ f(1, 3) &= \frac{3}{16}; & f(2, 3) &= \frac{1}{16} \end{aligned}$$

- (i) Determine the following probabilities: (a)  $P(X_1 \geq X_2)$ ; (b)  $P(X_1 + X_2 = 4)$ ; (c)  $P(|X_2 - X_1| = 1)$ ; (d)  $P(X_1 + X_2 \text{ is even})$ .  
(ii) Obtain the joint cumulative distribution function,  $F(x_1, x_2)$ .

**5.3** In a game of chess, one player either wins,  $W$ , loses,  $L$ , or draws,  $D$  (either by mutual agreement with the opponent, or as a result of a “stalemate”). Consider a player participating in a two-game, pre-tournament qualification series:

- (i) Obtain the sample space,  $\Omega$ .  
(ii) Define the two-dimensional random variable  $(X_1, X_2)$  where  $X_1$  is the total number of wins, and  $X_2$  is the total number of draws. Obtain  $V$  and, assuming equiprobable outcomes in the original sample space, determine the complete joint pdf,  $f(x_1, x_2)$ .  
(iii) If the player is awarded 3 points for a win, 1 point for a draw and no point for a loss, define the random variable  $Y$  as the total number of points assigned to a player

at the end of the two-game preliminary round. If a player needs at least 4 points to qualify, determine the probability of qualifying.

**5.4** Revisit Exercise 5.3 above but this time consider three players: Suzie, the superior player for whom the probability of winning a game,  $p_W = 0.75$ , the probability of drawing,  $p_D = 0.2$  and the probability of losing,  $p_L = 0.05$ ; Meredith, the mediocre player for whom  $p_W = 0.5$ ;  $p_D = 0.3$ ;  $p_L = 0.2$ ; and Paula, the poor player, for whom  $p_W = 0.2$ ;  $p_D = 0.3$ ;  $p_L = 0.5$ . Determine the complete joint pdf for each player,  $f_S(x_1, x_2)$ , for Suzie,  $f_M(x_1, x_2)$ , for Meredith, and  $f_P(x_1, x_2)$ , for Paula; and from these, determine for each player, the probability that she qualifies for the tournament.

**5.5** The continuous random variables  $X_1$  and  $X_2$  have the joint pdf

$$f(x, y) = \begin{cases} cx_1x_2(1 - x_2); & 0 < x_1 < 2; 0 < x_2 < 1 \\ 0; & \text{elsewhere} \end{cases} \quad (5.93)$$

- (i) Find the value of  $c$  if this is to be a valid pdf.
- (ii) Determine  $P(1 < x_1 < 2; 0.5 < x_2 < 1)$  and  $P(x_1 > 1; x_2 < 0.5)$ .
- (iii) Determine  $F(x_1, x_2)$ .

**5.6** Revisit Exercise 5.5.

- (i) Obtain the marginal pdfs  $f_1(x_1)$ ,  $f_2(x_2)$ , and the marginal means,  $\mu_{X_1}, \mu_{X_2}$ . Are  $X_1$  and  $X_2$  independent?
- (ii) Obtain the conditional pdf's  $f(x_1|x_2)$  and  $f(x_2|x_1)$ .

**5.7** The joint pdf  $f(x_1, x_2)$  for a two-dimensional random variable is given by the following table:

| $X_1 \rightarrow$ | 0     | 1     | 2     |
|-------------------|-------|-------|-------|
| $X_2 \downarrow$  | 0     | 0     | $1/4$ |
| 0                 | 0     | $1/2$ | 0     |
| 1                 | $1/4$ | 0     | 0     |
| 2                 |       |       |       |

- (i) Obtain the marginal pdfs,  $f_1(x_1)$  and  $f_2(x_2)$ , and determine whether or not  $X_1$  and  $X_2$  are independent.
- (ii) Obtain the conditional pdfs  $f(x_1|x_2)$  and  $f(x_2|x_1)$ . Describe in words what these results imply in terms of the original “experiments” and these random variables.
- (iii) It is conjectured that this joint pdf is for an experiment involving tossing a fair coin twice, with  $X_1$  as the total number of heads, and  $X_2$  as the total number of tails. Are the foregoing results consistent with this conjecture? Explain.

**5.8** Given the joint pdf:

$$f(x_1, x_2) = \begin{cases} ce^{-(x_1+x_2)}; & 0 < x_1 < 1; 0 < x_2 < 2; \\ 0; & \text{elsewhere} \end{cases} \quad (5.94)$$

First obtain  $c$ , then obtain the marginal pdfs  $f_1(x_1)$  and  $f_2(x_2)$ , and hence determine whether or not  $X_1$  and  $X_2$  are independent.

**5.9** If the range of validity of the joint pdf in Exercise 5.8 and Eq (5.94) are modified to  $0 < x_1 < \infty$  and  $0 < x_2 < \infty$ , obtain  $c$  and the marginal pdf, and then determine whether or not these random variables are now independent.

### Section 5.3

**5.10** Revisit Exercise 5.3. From the joint pdf determine

- (i)  $E[U(X_1, X_2)] = X_1 + X_2]$ .
- (ii)  $E[U(X_1, X_2)] = 3X_1 + X_2]$ . Use this result to determine if the player will be expected to qualify or not.

**5.11** For each of the three players in Exercise 5.4,

- (i) Determine the marginal pdfs,  $f_1(x_1)$  and  $f_2(x_2)$  and the marginal means  $\mu_{X_1}, \mu_{X_2}$ .
- (ii) Determine  $E[U(X_1, X_2)] = 3X_1 + X_2]$  and use the result to determine which of the three players, if any, will be expected to qualify for the tournament.

**5.12** Determine the covariance and correlation coefficient for the two random variables whose joint pdf,  $f(x_1, x_2)$  is given in the table in Exercise 5.7.

**5.13** For each of the three chess players in Exercise 5.4, Suzie, Meredith, and Paula, and from the joint pdf of each player's performance at the pre-tournament qualifying games, determine the covariance and correlation coefficients for each player. Discuss what these results imply in terms of the relationship between wins and draws for each player.

**5.14** The joint pdf for two random variables  $X$  and  $Y$  is given as:

$$f(x, y) = \begin{cases} x + y; & 0 < x < 1; 0 < y < 1; \\ 0; & \text{elsewhere} \end{cases} \quad (5.95)$$

- (i) Obtain  $f(x|y)$  and  $f(y|x)$  and show that these two random variables are *not* independent.
- (ii) Obtain the covariance,  $\sigma_{XY}$ , and the correlation coefficient,  $\rho$ . Comment on the strength of the correlation between these two random variables.

## APPLICATION PROBLEMS

**5.15** Refer to Application Problem 3.23 in Chapter 3, where the relationship between a blood assay used to determine lithium concentration in blood samples and lithium toxicity in 150 patients was presented in a table reproduced here for ease of reference.

| Assay | Lithium Toxicity |       | Total |
|-------|------------------|-------|-------|
|       | $L^+$            | $L^-$ |       |
| $A^+$ | 30               | 17    | 47    |
| $A^-$ | 21               | 82    | 103   |
| Total | 51               | 92    | 150   |

$A^+$  indicates high lithium concentrations in the blood assay and  $A^-$  indicates low lithium concentration;  $L^+$  indicates confirmed Lithium toxicity and  $L^-$  indicates *no* lithium toxicity.

- (i) In general, consider the assay result as the random variable  $Y$  having two possible outcomes  $y_1 = A^+$ , and  $y_2 = A^-$ ; and consider the true lithium toxicity status as the random variable  $X$  also having two possible outcomes  $x_1 = L^+$ , and  $x_2 = L^-$ . Now consider that the “relative frequencies” (or proportions) indicated in the data table can be approximately considered as close enough to true probabilities; convert the data table to a table of joint probability distribution  $f(x, y)$ . What is the probability that the test method will produce the right result?
- (ii) From the table of the joint pdf, compute the following probabilities and explain what they mean in words in terms of the problem at hand:  $f(y_2|x_2)$ ;  $f(y_1|x_2)$ ;  $f(y_2|x_1)$ .

**5.16** The reliability of the temperature control system for a commercial, highly exothermic polymer reactor presented in Example 5.2 in the text is known to depend on the lifetimes (in years) of the control hardware electronics,  $X_1$ , and of the control valve on the cooling water line,  $X_2$ ; the joint pdf is:

$$f(x_1, x_2) = \begin{cases} \frac{1}{50}e^{-(0.2x_1+0.1x_2)}; & 0 < x_1 < \infty \\ & 0 < x_2 < \infty \\ 0 & \text{elsewhere} \end{cases}$$

- (i) Determine the probability that the control valve outlasts the control hardware electronics.
- (ii) Determine the converse probability that the controller hardware electronics outlast the control valve.
- (iii) If a component is replaced every time it fails, how frequently can one expect to replace the control valve, and how frequently can one expect to replace the controller hardware electronics?
- (iv) If it costs \$20,000 to replace the control hardware electronics and \$10,000 to replace the control valve, how much should be budgeted over the next 20 years for keeping the control system functioning, assuming all other characteristics remain essentially the same over this period?

**5.17** In a major bio-vaccine research company, it is inevitable that workers are exposed to some hazardous, but highly treatable, disease causing agents. According to papers filed with the Safety and Hazards Authorities of the state in which the facility is located, the treatment provided is tailored to the worker’s age, (the variable,  $X$ : 0 if younger than 30 years; 1 if 31 years or older), and location in the facility (a surrogate for virulence of the proprietary strains used in various parts of the facility, represented by the variable  $Y = 1, 2, 3$  or 4. The composition of the 2,500 employees at the company’s research headquarters is shown in the table below:

| Location → | 1   | 2   | 3   | 4   |
|------------|-----|-----|-----|-----|
| Age ↓      |     |     |     |     |
| $< 30$     | 6%  | 20% | 13% | 10% |
| $\geq 31$  | 17% | 14% | 12% | 8%  |

- (i) If a worker is infected at random so that the outcome is the bivariate random variable  $(X, Y)$  where  $X$  has two outcomes, and  $Y$  has four, obtain the pdf  $f(x, y)$  from the given data (assuming each worker in each location has an equal chance of infection); and determine the marginal pdf’s  $f_1(x)$  and  $f_2(y)$ .

- (ii) What is the probability that a worker in need of treatment was infected in location 3 or 4 given that he/she is < 30 years old?
- (iii) If the cost of the treating each infected worker (in dollars per year) is given by the expression

$$C = 1500 - 100Y + 500X \quad (5.96)$$

how much should the company expect to spend per worker every year, assuming the worker composition remains the same year after year?

**5.18** A non-destructive quality control test on a military weapon system correctly detects a flaw in the central electronic guidance subunit if one exists, *or* correctly accepts the system as fully functional if no flaw exists, 85% of the time; it incorrectly identifies a flaw when one does not exist (a false positive), 5% of the time, and incorrectly fails to detect a flaw when one exists (a false negative), 10% of the time. When the test is repeated 5 times under mostly identical conditions, if  $X_1$  is the number of times the test is correct, and  $X_2$  is the number of times it registers a false positive, the joint pdf of these two random variables is given as:

$$f(x_1, x_2) = \frac{120}{x_1!x_2!} 0.85^{x_1} 0.05^{x_2} \quad (5.97)$$

- (i) Why is no consideration given in the expression in Eq (5.97) to the third random variable,  $X_3$ , the number of times the test registers a false negative?
- (ii) From Eq (5.97), generate a  $5 \times 5$  table of  $f(x_1, x_2)$  for all the possible outcomes and from this obtain the marginal pdfs,  $f_1(x_1)$  and  $f_2(x_2)$ . Are these two random variables independent?
- (iii) Determine the expected number of correct test results regardless of the other results; also determine the expected value of false positives regardless of other results.
- (iv) What is the expected number of the total number of correct results *and* false positives? Is this value the same as the sum of the expected values obtained in (iii)? Explain.

# Chapter 6

## Random Variable Transformations

|       |   |     |
|-------|---|-----|
| 6.1   | Introduction and Problem Definition .....           | 171 |
| 6.2   | Single Variable Transformations .....               | 172 |
| 6.2.1 | Discrete Case .....                                 | 173 |
|       | A Practical Application .....                       | 173 |
| 6.2.2 | Continuous Case .....                               | 175 |
| 6.2.3 | General Continuous Case .....                       | 176 |
| 6.2.4 | Random Variable Sums .....                          | 177 |
|       | The Cumulative Distribution Function Approach ..... | 177 |
|       | The Characteristic Function Approach .....          | 179 |
| 6.3   | Bivariate Transformations .....                     | 181 |
| 6.4   | General Multivariate Transformations .....          | 184 |
| 6.4.1 | Square Transformations .....                        | 184 |
| 6.4.2 | Non-Square Transformations .....                    | 185 |
| 6.4.3 | Non-Monotone Transformations .....                  | 188 |
| 6.5   | Summary and Conclusions .....                       | 188 |
|       | REVIEW QUESTIONS .....                              | 189 |
|       | EXERCISES .....                                     | 190 |
|       | APPLICATION PROBLEMS .....                          | 192 |

*From a god to a bull! a heavy descension!  
it was Jove's case.  
From a prince to a prentice!  
a low transformation!  
that shall be mine; for in every thing  
the purpose must weigh with the folly. Follow me, Ned.*

*King Henry the Fourth,  
William Shakespeare (1564–1616)*

Many problems of practical interest involve a random variable  $Y$  that is defined as a function of another random variable  $X$ , say according to  $Y = \phi(X)$ , so that the characteristics of the one arise directly from those of the other via the indicated transformation. In particular, if we already know the probability distribution function for  $X$  as  $f_X(x)$ , it will be helpful to know how to determine the corresponding distribution function for  $Y$ . This chapter presents techniques for characterizing functions of random variables, and the results, important in their own right, become particularly useful in Part III where probability models are derived for random phenomena of importance in engineering and science.

## 6.1 Introduction and Problem Definition

The problem of primary interest to us in this chapter may be stated as follows:

Given a random variable  $X$  with pdf  $f_X(x)$ , we are interested in deriving an expression for the corresponding pdf  $f_Y(y)$  for the random variable  $Y$  related to  $X$  according to:

$$Y = \phi(X) \quad (6.1)$$

More generally, given the  $n$ -dimensional random variable  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  with the joint pdf  $f_{\mathbf{X}}(\mathbf{x})$ , we want to find the corresponding pdf  $f_{\mathbf{Y}}(\mathbf{y})$ , for the  $m$ -dimensional random variable  $\mathbf{Y} = (Y_1, Y_2, \dots, Y_m)$  when the two are related according to:

$$\begin{aligned} Y_1 &= \phi_1(X_1, X_2, \dots, X_n); \\ Y_2 &= \phi_2(X_1, X_2, \dots, X_n); \\ \dots &= \dots \\ Y_m &= \phi_m(X_1, X_2, \dots, X_n) \end{aligned} \quad (6.2)$$

As demonstrated in later chapters, these results are extremely useful in deriving probability models for more complicated random variables from the probability models of simpler ones.

## 6.2 Single Variable Transformations

We begin with the simplest case when  $Y$  is a function of a single variable  $X$

$$Y = \phi(X); \forall X \in V_X \quad (6.3)$$

$\phi$  is a continuous function that transforms each point  $x$  in  $V_X$ , the space over which the random variable  $X$  is defined, to  $y$ , thereby mapping  $V_X$  onto the corresponding space  $V_Y$  for the resulting random variable  $Y$ . Furthermore, this transformation is one-to-one in the sense that each point in  $V_X$  corresponds to one and only one point in  $V_Y$ . In this case, the inverse transformation,

$$X = \psi(Y); \forall Y \in V_Y \quad (6.4)$$

exists and is also one-to-one. The procedure for obtaining  $f_Y(y)$  given  $f_X(x)$  is highly dependent on the nature of the random variable in question, being more straightforward for the discrete case than for the continuous.

### 6.2.1 Discrete Case

When  $X$  is a discrete random variable, we have

$$f_Y(y) = P(Y = y) = P(X = \psi(y)) = f_X[\psi(y)]; \forall Y \in V_Y \quad (6.5)$$

We illustrate this straightforward result first with the following simple example.

#### Example 6.1 LINEAR TRANSFORMATION OF A POISSON RANDOM VARIABLE

As discussed in more detail in Part III, the discrete random variable  $X$  having the following pdf:

$$f_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}; x = 0, 1, 2, 3, \dots \quad (6.6)$$

is a Poisson random variable; it provides a useful model of random phenomena involving the occurrence of rare events in a finite interval of length, time or space. Find the pdf  $f_Y(y)$  for the random variable  $Y$  related to  $X$  according to:

$$Y = 2X. \quad (6.7)$$

#### Solution:

First we note that the transformation in Eq (6.7) is one-to-one, mapping  $V_X = \{0, 1, 2, 3, \dots\}$  onto  $V_Y = \{0, 2, 4, 6, \dots\}$ ; the inverse transformation is:

$$X = \frac{1}{2}Y \quad (6.8)$$

so that from Eq (6.5) we obtain:

$$\begin{aligned} f_Y(y) &= P(Y = y) = P(X = y/2) \\ &= \frac{\lambda^{y/2} e^{-\lambda}}{(y/2)!}; y = 0, 2, 4, 6, \dots \end{aligned} \quad (6.9)$$

Thus, under the transformation  $Y = 2X$  and  $f_X(x)$  as given in Eq (6.6), the desired pdf  $f_Y(y)$  is given by:

$$f_Y(y) = \frac{\lambda^{y/2} e^{-\lambda}}{(y/2)!}; y = 0, 2, 4, 6, \dots \quad (6.10)$$

### A Practical Application

The number of times,  $X$ , that each cell in a cell culture divides in a time interval of length,  $t$ , is a random variable whose specific value depends on many factors both intrinsic (e.g. individual cell characteristics) and extrinsic

(e.g. media characteristics, temperature, oxygen). As discussed in Chapter 8, the underlying random phenomenon matches well to that of the ideal Poisson random variable, so that if  $\eta$  is the mean rate of division per unit time associated with a particular cell population, the probability distribution of  $X$  is given by Eq (6.11)

$$f_X(x) = \frac{(\eta t)^x e^{-(\eta t)}}{x!}; x = 0, 1, 2, 3, \dots \quad (6.11)$$

where, in terms of Eq (6.6),

$$\lambda = \eta t \quad (6.12)$$

In many cases, however, the cell culture characteristic of interest is not so much the *number of times* that each cell divides as it is the *number of cells*,  $Y$ , in the culture after the passage of a specific amount of time. For each cell in the culture, the relationship between these two random variables is given by:

$$Y = 2^X \quad (6.13)$$

The problem of interest is now to find  $f_Y(y)$ , the pdf of the number of cells in the culture, given  $f_X(x)$ .

As with the simple example given above, note that the transformation in (6.13), even though nonlinear, is one-to-one, mapping  $V_X = \{0, 1, 2, 3, \dots\}$  onto  $V_Y = \{1, 2, 4, 8, \dots\}$ ; the inverse transformation is:

$$X = \log_2 Y \quad (6.14)$$

From here, we easily obtain the required  $f_Y(y)$  as:

$$f_Y(y) = \frac{e^{-\lambda} \lambda^{\log_2 y}}{(\log_2 y)!}; y = 1, 2, 4, 8, \dots \quad (6.15)$$

a somewhat unwieldy-looking, but nonetheless valid, pdf that can be simplified a bit by noting that:

$$\lambda^{\log_2 y} = y^{\log_2 \lambda} = y^\theta \quad (6.16)$$

if we define

$$\theta = \log_2 \lambda \quad (6.17)$$

a logarithmic transformation of the Poisson parameter  $\lambda$ . Thus:

$$f_Y(y) = \frac{e^{-(2^\theta)} y^\theta}{(\log_2 y)!}; y = 1, 2, 4, 8, \dots \quad (6.18)$$

It is possible to confirm that the pdf obtained in Eq (6.18) for  $Y$ , the number of cells in the culture after a time interval  $t$  is a valid pdf for which:

$$\sum_y f_Y(y) = 1 \quad (6.19)$$

since, from Eq (6.18)

$$\begin{aligned}\sum_y f_Y(y) &= e^{-(2^\theta)} \left[ 1 + \frac{2^\theta}{1} + \frac{2^{2\theta}}{2!} + \frac{2^{3\theta}}{3!} + \dots \right] \\ &= e^{-(2^\theta)} e^{(2^\theta)} = 1\end{aligned}\quad (6.20)$$

The mean number of cells in the culture after time  $t$ ,  $E[Y]$ , can be shown (see end-of-chapter Exercise 6.2) to be:

$$E[Y] = e^\lambda \quad (6.21)$$

which should be compared with  $E[X] = \lambda$ .

### 6.2.2 Continuous Case

When  $X$  is a continuous random variable, things are slightly different. In addition to the inverse transformation given in Eq (23.50), let us define the function:

$$J = \frac{d}{dy}[\psi(y)] = \psi'(y) \quad (6.22)$$

known as the *Jacobian of the inverse transformation*. If the transformation is such that it is strictly monotonic (increasing or decreasing), then it can be shown that:

$$f_Y(y) = f_X[\psi(y)]|J|; \forall Y \in V_Y \quad (6.23)$$

The argument goes as follows: If  $F_Y(y)$  is the cdf for the new variable  $Y$ , then

$$F_Y(y) = P(Y \leq y) = P(\phi(X) \leq y) = P(X \leq \psi(y)) = F_X[\psi(y)] \quad (6.24)$$

and by differentiation, we obtain:

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{d}{dy} \{F_X[\psi(y)]\} = f_X[\psi(y)] \frac{d}{dy} \{\psi(y)\} \quad (6.25)$$

with the derivative on the RHS positive for a strictly monotonic increasing function. It can be shown that if  $\phi$  were monotonically decreasing, the expression in (6.24) will yield:

$$f_Y(y) = -f_X[\psi(y)] \frac{d}{dy} \{\psi(y)\} \quad (6.26)$$

with the derivative on the RHS as a negative quantity. Both results may be combined into one as

$$f_Y(y) = f_X[\psi(y)] \left| \frac{d}{dy} \{\psi(y)\} \right| \quad (6.27)$$

as presented in Eq (6.23). Let us illustrate this with another example.

**Example 6.2 LOG TRANSFORMATION OF A UNIFORM RANDOM VARIABLE**

The random variable  $X$  with the following pdf:

$$f_X(x) = \begin{cases} 1; & 0 < x < 1 \\ 0; & \text{otherwise} \end{cases} \quad (6.28)$$

is identified in Part III as the *uniform* random variable. Determine the pdf for the random variable  $Y$  obtained via the transformation:

$$Y = -\beta \ln X \quad (6.29)$$

**Solution:**

The transformation is one-to-one, maps  $V_X = \{0 < x < 1\}$  onto  $V_Y = \{0 < y < \infty\}$ , and the inverse transformation is given by:

$$X = \psi(y) = e^{-y/\beta}; 0 < y < \infty. \quad (6.30)$$

The Jacobian of the inverse transformation is:

$$J = \psi'(y) = -\frac{1}{\beta}e^{-y/\beta} \quad (6.31)$$

Thus, from Eq (6.23) or Eq (6.27), we obtain the required pdf as:

$$f_Y(y) = f_X[\psi(y)]|J| = \begin{cases} \frac{1}{\beta}e^{-y/\beta}; & 0 < y < \infty \\ 0; & \text{otherwise} \end{cases} \quad (6.32)$$

These two random variables and their corresponding models are discussed more fully in Part III.

### 6.2.3 General Continuous Case

When the transformation  $Y = \phi(X)$  is not strictly monotone, the result given above is modified as follows: Let the function  $y = \phi(x)$  possess a countable number of roots,  $x_i$ , represented as a function of  $y$  as:

$$x_i = \phi_i^{-1}(y) = \psi_i(y); i = 1, 2, 3, \dots, k \quad (6.33)$$

with corresponding Jacobians:

$$J_i = \frac{d}{dy}\{\psi_i(y)\} \quad (6.34)$$

then it can be shown that Eq (6.23) (or equivalently Eq (6.27)) becomes:

$$f_Y(y) = \sum_{i=1}^k f_X[\psi_i(y)]|J_i| \quad (6.35)$$

Let us illustrate with an example.

**Example 6.3 THE SQUARE OF A STANDARD NORMAL RANDOM VARIABLE**

The random variable  $X$  has the following pdf:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}; -\infty < x < \infty \quad (6.36)$$

Determine the pdf for the random variable  $Y$  obtained via the transformation:

$$y = x^2 \quad (6.37)$$

**Solution:**

Observe that this transformation, which maps the space  $V_X = -\infty < x < \infty$  onto  $V_Y = 0 < y < \infty$ , is *not* one-to-one; for all  $y > 0$  there are two  $x$ 's corresponding to each  $y$ , since the inverse transformation is given by:

$$x = \pm\sqrt{y} \quad (6.38)$$

The transformation thus has 2 roots for  $x$ :

$$\begin{aligned} x_1 &= \psi_1(y) = \sqrt{y} \\ x_2 &= \psi_2(y) = -\sqrt{y} \end{aligned} \quad (6.39)$$

and upon computing the corresponding derivatives, Eq (6.35) becomes

$$f_Y(y) = f_X(\sqrt{y}) \frac{y^{-1/2}}{2} + f_X(-\sqrt{y}) \frac{y^{-1/2}}{2} \quad (6.40)$$

which simplifies to:

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-y/2} y^{-1/2}; 0 < y < \infty \quad (6.41)$$

This important result is used later.

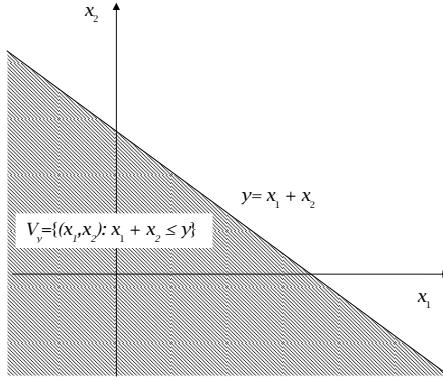
#### 6.2.4 Random Variable Sums

Let us consider first the case where the random variable transformation involves the sum of two independent random variables, i.e.,

$$Y = \phi(X_1, X_2) = X_1 + X_2 \quad (6.42)$$

where  $f_1(x_1)$  and  $f_2(x_2)$ , are, respectively, the known pdfs of the random variables  $X_1$  and  $X_2$ . Two approaches are typically employed in finding the desired  $f_Y(y)$ :

- The cumulative distribution function approach;
- The characteristic function approach.



**FIGURE 6.1:** Region of interest,  $V_Y$ , for computing the cdf of the random variable  $Y$  defined as a sum of 2 independent random variables  $X_1$  and  $X_2$

### The Cumulative Distribution Function Approach

This approach requires first obtaining the cdf  $F_Y(y)$  (as argued in Eq (6.24)), from where the desired pdf is obtained by differentiation when  $Y$  is continuous. In this case, the cdf  $F_Y(y)$  is obtained as:

$$F_Y(y) = P(Y \leq y) = \int \int_{V_Y} f(x_1, x_2) dx_1 dx_2 \quad (6.43)$$

where  $f(x_1, x_2)$  is the joint pdf of  $X_1$  and  $X_2$ , and, most importantly, the region over which the double integration is being carried out,  $V_Y$ , is given by:

$$V_Y = \{(x_1, x_2) : x_1 + x_2 \leq y\} \quad (6.44)$$

as shown in Fig 6.1. Observe from this figure that the integration may be carried out several different ways: if we integrate first with respect to  $x_1$ , the limits go from  $-\infty$  until we reach the line, at which point  $x_1 = y - x_2$ ; we then integrate with respect to  $x_2$  from  $-\infty$  to  $\infty$ . In this case, Eq (6.43) becomes:

$$F_Y(y) = \int_{-\infty}^{\infty} \int_{-\infty}^{y-x_2} f(x_1, x_2) dx_1 dx_2 \quad (6.45)$$

from where we may differentiate with respect to  $y$  to obtain:

$$f_Y(y) = \int_{-\infty}^{\infty} f(y - x_2, x_2) dx_2 \quad (6.46)$$

In particular, if  $X_1$  and  $X_2$  are independent so that the joint pdf is a product of the individual marginal pdfs, we obtain:

$$f_Y(y) = \int_{-\infty}^{\infty} f_1(y - x_2) f_2(x_2) dx_2 \quad (6.47)$$

If, instead, the integration in Eq (6.43) had been done first with respect to  $x_2$  and then with respect to  $x_1$ , the resulting differentiation would have resulted in the alternative, and entirely equivalent, expression:

$$f_Y(y) = \int_{-\infty}^{\infty} f_2(y - x_1) f_1(x_1) dx_1 \quad (6.48)$$

Integrals of this nature are known as *convolutions* of the functions  $f_1(x_1)$  and  $f_2(x_2)$  and this is as far as we can go with a general discussion.

Thus, we have the general result that the pdf of the random variable  $Y$  obtained as a sum of two independent random variables  $X_1$  and  $X_2$  is a convolution of the two contributing pdfs  $f_1(x_1)$  and  $f_2(x_2)$  as shown in Eqs (6.47) and (6.48).

Let us illustrate this with a classic example.

#### Example 6.4 THE SUM OF TWO EXPONENTIAL RANDOM VARIABLES

Given two stochastically independent random variables  $X_1$  and  $X_2$  with pdf's:

$$f_1(x_1) = \frac{1}{\beta} e^{-x_1/\beta}; 0 < x_1 < \infty \quad (6.49)$$

$$f_2(x_2) = \frac{1}{\beta} e^{-x_2/\beta}; 0 < x_2 < \infty \quad (6.50)$$

Determine the pdf of the random variable  $Y = X_1 + X_2$ .

**Solution:**

In this case, the required pdf is obtained from the convolution:

$$f_Y(y) = \frac{1}{\beta^2} \int_{-\infty}^{\infty} e^{-(y-x_2)/\beta} e^{-x_2/\beta} dx_2; 0 < y < \infty \quad (6.51)$$

However, because  $x_2$  is non-negative, as  $x_1 = y - x_2$  must also be, the limits on the integral have to be restricted to go from  $x_2 = 0$  to  $x_2 = y$ ; so that:

$$f_Y(y) = \frac{1}{\beta^2} \int_0^y e^{-(y-x_2)/\beta} e^{-x_2/\beta} dx_2; 0 < y < \infty \quad (6.52)$$

Upon carrying out the indicated integral, we obtain the final result:

$$f_Y(y) = \frac{1}{\beta^2} y e^{-y/\beta}; 0 < y < \infty \quad (6.53)$$

Observe that the result presented above for the sum of two random variables extends directly to the sum of more than two random variables by successive additions. However, this procedure becomes rapidly more tedious as we must carry out repeated convolution integrals over increasingly more complex regions.

### The Characteristic Function Approach

It is far more convenient to employ the characteristic function to determine the pdf of random variable sums, continuous or discrete. The pertinent result is from a property discussed earlier in Chapter 4: for independent random variables  $X_1$  and  $X_2$  with respective characteristic functions  $\varphi_{X_1}(t)$  and  $\varphi_{X_2}(t)$ , the characteristic function of their sum  $Y = X_1 + X_2$  is given by:

$$\varphi_Y(t) = \varphi_{X_1}(t)\varphi_{X_2}(t) \quad (6.54)$$

In general, for  $n$  independent random variables  $X_i; i = 1, 2, \dots, n$ , each with respective characteristic functions,  $\varphi_{X_i}(t)$ , if

$$Y = X_1 + X_2 + \dots + X_n \quad (6.55)$$

then

$$\varphi_Y(t) = \varphi_{X_1}(t)\varphi_{X_2}(t) \cdots \varphi_{X_n}(t) \quad (6.56)$$

The utility of this result lies in the fact that  $\varphi_Y(t)$  is easily obtained from each contributing  $\varphi_{X_i}(t)$ ; the desired  $f_Y(y)$  is then recovered from  $\varphi_Y(t)$  either by inspection (when this is obvious), or else by the inversion formula presented in Chapter 4.

Let us illustrate this with the same example used above.

#### Example 6.5 THE SUM OF TWO EXPONENTIAL RANDOM VARIABLES REVISITED

Using characteristic functions, determine the pdf of the random variable  $Y = X_1 + X_2$ , where the pdfs of the two stochastically independent random variables  $X_1$  and  $X_2$  are as given in Example 6.4 above and their characteristic functions are given as:

$$\varphi_{X_1}(t) = \varphi_{X_2}(t) = \frac{1}{(1 - j\beta t)} \quad (6.57)$$

#### Solution:

From Eq (6.54), the required characteristic function for the sum is:

$$\varphi_Y(t) = \frac{1}{(1 - j\beta t)^2} \quad (6.58)$$

At this point, anyone familiar with specific random variable pdfs and their characteristic functions will recognize this particular form right away: it is the pdf of a gamma random variable, specifically  $\gamma(2, \beta)$ , as Chapter 9 shows. However, since we have not yet introduced these important random variables, their pdfs and characteristic functions (see Chapter 9), we therefore do not expect the reader to be able to deduce the pdf corresponding to  $\varphi_Y(t)$  above by inspection. In this case we can invoke the inversion formula of Chapter 4 to obtain:

$$\begin{aligned} f_Y(y) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-jyt} \varphi_Y(t) dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-jyt}}{(1 - j\beta t)^2} dt \end{aligned} \quad (6.59)$$

Upon carrying out the indicated integral, we obtain the final result:

$$f_Y(y) = \frac{1}{\beta^2} y e^{-y/\beta}; 0 < y < \infty \quad (6.60)$$

In general, it is not necessary to carry out the inversion integration explicitly once one becomes familiar with characteristic functions of various pdfs. (To engineers familiar with the application of Laplace transforms to the solution of ordinary differential equations, this is identical to how tables of inverse transforms have eliminated the need for explicitly carrying out Laplace inversions.) This point is illustrated in the next anticipatory example (and in subsequent chapters).

**Example 6.6 REPRODUCTIVE PROPERTY OF GAMMA RANDOM VARIABLE**

A random variable,  $X$ , with the following pdf

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} e^{-x/\beta} x^{\alpha-1}; 0 < x < \infty \quad (6.61)$$

is identified in Chapter 9 as a gamma random variable with parameters  $\alpha$  and  $\beta$ . Its characteristic function is:

$$\varphi_X(t) = \frac{1}{(1 - j\beta t)^\alpha} \quad (6.62)$$

Find the pdf of the random variable  $Y$  defined as the sum of the  $n$  independent such random variables,  $X_i$ , each with different parameters  $\alpha_i$  but with the same parameter  $\beta$ .

**Solution:**

The desired transformation is

$$Y = \sum_{i=1}^n X_i \quad (6.63)$$

and from the given individual characteristic functions for each  $X_i$ , we obtain the required characteristic function for the sum  $Y$  as:

$$\varphi_Y(t) = \prod_{i=1}^n \varphi_{X_i}(t) = \frac{1}{(1 - j\beta t)^{\alpha^*}} \quad (6.64)$$

where  $\alpha^* = \sum_{i=1}^n \alpha_i$ . Now, by comparing Eq (6.62) with Eq (6.64), we see immediately the important result that  $Y$  is *also* a gamma random variable, with parameters  $\alpha^*$  and  $\beta$ . Thus, this sum of gamma random variables begets another gamma random variable, a result generally known as the “reproductive property” of the gamma random variable.

### 6.3 Bivariate Transformations

Because of its many practical applications, it is instructive to consider first the bivariate case before taking on the full multivariate problem. In this case we are concerned with determining the joint pdf  $f_{\mathbf{Y}}(\mathbf{y})$  for the 2-dimensional random variable  $\mathbf{Y} = (Y_1, Y_2)$  obtained from the 2-dimensional random variable  $\mathbf{X} = (X_1, X_2)$  with the known joint pdf  $f_{\mathbf{X}}(\mathbf{x})$ , via the following bivariate transformation:

$$\begin{aligned} Y_1 &= \phi_1(X_1, X_2) \\ Y_2 &= \phi_2(X_1, X_2) \end{aligned} \quad (6.65)$$

written more compactly as:

$$\mathbf{Y} = \Phi(\mathbf{X}) \quad (6.66)$$

As in the single variable case, we consider first the case where these functions are continuous and collectively define a one-to-one transformation that maps the two-dimensional space  $V_X$  in the  $x_1 - x_2$  plane to the two-dimensional space  $V_Y$  in the  $y_1 - y_2$  plane. The inverse bivariate transformation is given by:

$$\begin{aligned} X_1 &= \psi_1(Y_1, Y_2) \\ X_2 &= \psi_2(Y_1, Y_2) \end{aligned} \quad (6.67)$$

or, more compactly,

$$\mathbf{X} = \Psi(\mathbf{Y}) \quad (6.68)$$

The  $2 \times 2$  determinant given by:

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix} \quad (6.69)$$

is the Jacobian of this bivariate inverse transformation, and so long as  $J$  does not vanish identically in  $V_Y$ , it can be shown that the desired joint pdf for  $\mathbf{Y}$  is given by:

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}[\Psi(\mathbf{y})] |J|; \forall \mathbf{y} \in V_Y \quad (6.70)$$

where the similarity with Eq (6.27) should not be lost on the reader. The following is a classic example typically used to illustrate this result.

#### Example 6.7 RELATING GAMMA AND BETA RANDOM VARIABLES

Given two stochastically independent random variables  $X_1$  and  $X_2$  with pdfs:

$$f_1(x_1) = \frac{1}{\Gamma(\alpha)} x_1^{\alpha-1} e^{-x_1}; 0 < x_1 < \infty \quad (6.71)$$

$$f_2(x_2) = \frac{1}{\Gamma(\beta)} x_1^{\beta-1} e^{-x_2}; 0 < x_2 < \infty \quad (6.72)$$

Determine *both* the joint and the marginal pdfs for the two random variables  $Y_1$  and  $Y_2$  obtained via the transformation:

$$\begin{aligned} Y_1 &= X_1 + X_2 \\ Y_2 &= \frac{X_1}{X_1 + X_2} \end{aligned} \quad (6.73)$$

**Solution:**

First, by independence, the joint pdf for  $X_1$  and  $X_2$  is:

$$f_{\mathbf{X}}(x_1, x_2) = \frac{1}{\Gamma(\alpha)\Gamma(\beta)} x_1^{\alpha-1} x_2^{\beta-1} e^{-x_1} e^{-x_2}; 0 < x_1 < \infty; 0 < x_2 < \infty \quad (6.74)$$

Next, observe that the transformation in Eq (6.73) is a one-to-one mapping of  $V_X$ , the positive quadrant of the  $x_1 - x_2$  plane, onto  $V_Y = \{(y_1, y_2); 0 < y_1 < \infty, 0 < y_2 < 1\}$ ; the inverse transformation is given by:

$$\begin{aligned} x_1 &= y_1 y_2 \\ x_2 &= y_1(1 - y_2) \end{aligned} \quad (6.75)$$

and the Jacobian is obtained as

$$J = \begin{vmatrix} y_2 & y_1 \\ 1 - y_2 & -y_1 \end{vmatrix} = -y_1 \quad (6.76)$$

It vanishes at the point  $y_1 = 0$ , but this is a point of probability measure 0 that can be safely excluded from the space  $V_Y$ . Thus, from Eq (14.32), the joint pdf for  $Y_1$  and  $Y_2$  is:

$$f_{\mathbf{Y}}(y_1, y_2) = \begin{cases} \frac{1}{\Gamma(\alpha)\Gamma(\beta)} (y_1 y_2)^{\alpha-1} [y_1(1 - y_2)]^{\beta-1} e^{-y_1} y_1; & 0 < y_1 < \infty; \\ & 0 < y_2 < 1; \\ & \text{otherwise} \\ 0 & \end{cases} \quad (6.77)$$

This may be rearranged to give:

$$f_{\mathbf{Y}}(y_1, y_2) = \begin{cases} \frac{1}{\Gamma(\alpha)\Gamma(\beta)} \{y_2^{\alpha-1} (1 - y_2)^{\beta-1}\} \{e^{-y_1} y_1^{\alpha+\beta-1}\}; & 0 < y_1 < \infty; \\ & 0 < y_2 < 1; \\ & \text{otherwise} \\ 0 & \end{cases} \quad (6.78)$$

an equation which, apart from the constant, factors out into separate and distinct functions of  $y_1$  and  $y_2$ , indicating that the random variables  $Y_1$  and  $Y_2$  are independent.

By definition, the marginal pdf for  $Y_2$  is obtained by integrating out  $y_1$  in Eq (6.78) to obtain

$$f_2(y_2) = \frac{1}{\Gamma(\alpha)\Gamma(\beta)} y_2^{\alpha-1} (1 - y_2)^{\beta-1} \int_0^\infty e^{-y_1} y_1^{\alpha+\beta-1} dy_1 \quad (6.79)$$

Recognizing the integral as the gamma function, i.e.,

$$\Gamma(a) = \int_0^\infty e^{-y} y^{a-1} dy \quad (6.80)$$

we obtain:

$$f_2(y_2) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y_2^{\alpha-1} (1-y_2)^{\beta-1}; 0 < y_2 < 1 \quad (6.81)$$

Since, by independence,

$$f_{\mathbf{Y}}(y_1, y_2) = f_1(y_1)f_2(y_2) \quad (6.82)$$

it follows from Eqs (6.78), (6.71) or (6.72), and Eq (15.82) that the marginal pdf for  $Y_1$  is given by:

$$f_1(y_1) = \frac{1}{\Gamma(\alpha + \beta)} e^{-y_1} y_1^{\alpha+\beta-1}; 0 < y_1 < \infty \quad (6.83)$$

Again, we refer to these results later in Part III.

## 6.4 General Multivariate Transformations

As introduced briefly earlier, the general multivariate case is concerned with determining the joint pdf  $f_{\mathbf{Y}}(\mathbf{y})$  for the  $m$ -dimensional random variable  $\mathbf{Y} = (Y_1, Y_2, \dots, Y_m)$  arising from a transformation of the  $n$ -dimensional random variable  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  according to:

$$\begin{aligned} Y_1 &= \phi_1(X_1, X_2, \dots, X_n); \\ Y_2 &= \phi_2(X_1, X_2, \dots, X_n); \\ \dots &= \dots \\ Y_m &= \phi_m(X_1, X_2, \dots, X_n) \end{aligned}$$

given the joint pdf  $f_{\mathbf{X}}(\mathbf{x})$ .

### 6.4.1 Square Transformations

When  $n = m$  and the transformation is one-to-one, and the inverse transformation:

$$\begin{aligned} x_1 &= \psi_1(y_1, y_2, \dots, y_n); \\ x_2 &= \psi_2(y_1, y_2, \dots, y_n); \\ \dots &= \dots \\ x_n &= \psi_n(y_1, y_2, \dots, y_n) \end{aligned} \quad (6.84)$$

or, more compactly,

$$\mathbf{X} = \Psi(\mathbf{Y}) \quad (6.85)$$

yields the square  $n \times n$  determinant:

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{vmatrix} \quad (6.86)$$

And now, as in the bivariate case, it can be shown that for a  $J$  that is non-zero anywhere in  $V_Y$ , the desired joint pdf for  $Y$  is given by:

$$f_Y(\mathbf{y}) = f_X[\Psi(\mathbf{y})] |J|; \forall \mathbf{y} \in V_Y \quad (6.87)$$

an expression that is identical in every way to Eq (14.32) except for the dimensionality, and similar to the single variate result in Eq (6.23). Thus for the “square” transformation in which  $n = m$ , the required result is a direct generalization of the bivariate result, identical in structure, differing only in dimensionality.

#### 6.4.2 Non-Square Transformations

The case with  $n \neq m$  presents two different problems:

1.  $n < m$ ; the “overdefined” transformation in which there are more new variables  $\mathbf{Y}$  than the original  $\mathbf{X}$  variables;
2.  $n > m$ ; the “underdefined” transformation in which there are fewer new variables  $\mathbf{Y}$  than the original  $\mathbf{X}$  variables.

In the “overdefined” problem, it should be easy to see that there can be no exact inverse transformation except under some special, very restrictive circumstances, in which the extra  $(m - n)$   $\mathbf{Y}$  variables are merely redundant and can be expressed as functions of the other  $n$ . This problem is therefore of no practical interest: the general case has no exact solution; the special case reverts to the already solved square  $n \times n$  problem.

With the “underdefined” problem — the more common of the two — the strategy is to augment the  $m$  equations with an additional  $(m - n)$ , usually simple, variable transformations chosen such that an inverse transformation exists. Having thus “squared” the problem, the result in Eq (6.87) may then be applied to obtain a joint pdf for the augmented  $\mathbf{Y}$  variables. The final step involves integrating out the extraneous variables. This is best illustrated with some examples.

##### Example 6.8 SUM OF TWO STANDARD NORMAL RANDOM VARIABLES

Given two stochastically independent random variables  $X_1$  and  $X_2$  with pdfs:

$$f_1(x_1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x_1^2}{2}}; -\infty < x_1 < \infty \quad (6.88)$$

$$f_2(x_2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x_2^2}{2}}; -\infty < x_2 < \infty \quad (6.89)$$

determine the pdf of the random variable  $Y$  obtained from their sum,

$$Y = X_1 + X_2 \quad (6.90)$$

**Solution:**

First, observe that even though this is a sum, so that we could invoke earlier results to handle this problem, Eq (6.90) is also an underdetermined transformation from two dimensions in  $X_1$  and  $X_2$  to one in  $Y$ . To “square” the transformation, let the variable in Eq (6.90) now be  $Y_1$  and add another one, say  $Y_2 = X_1 - X_2$ , to give:

$$\begin{aligned} Y_1 &= X_1 + X_2 \\ Y_2 &= X_1 - X_2 \end{aligned} \quad (6.91)$$

which is now square, and one-to-one. The inverse transformation is:

$$\begin{aligned} x_1 &= \frac{1}{2}(y_1 + y_2) \\ x_2 &= \frac{1}{2}(y_1 - y_2) \end{aligned} \quad (6.92)$$

and a Jacobian,  $J = -1/2$ .

By independence, the joint pdf for  $X_1$  and  $X_2$  is given by:

$$f_{\mathbf{X}}(x_1, x_2) = \frac{1}{2\pi} e^{-\left(\frac{x_1^2+x_2^2}{2}\right)}; -\infty < x_1 < \infty; -\infty < x_2 < \infty \quad (6.93)$$

and from Eq (6.87), the joint pdf for  $Y_1$  and  $Y_2$  is obtained as:

$$f_{\mathbf{Y}}(y_1, y_2) = \frac{1}{2\pi} \frac{1}{2} e^{-\left(\frac{(y_1+y_2)^2+(y_1-y_2)^2}{8}\right)}; -\infty < y_1 < \infty; -\infty < y_2 < \infty \quad (6.94)$$

which rearranges easily to:

$$f_{\mathbf{Y}}(y_1, y_2) = \frac{1}{4\pi} e^{-\left(\frac{y_1^2}{4}\right)} e^{-\left(\frac{y_2^2}{4}\right)}; -\infty < y_1 < \infty; -\infty < y_2 < \infty \quad (6.95)$$

And now, either by inspection (this is a product of two clearly identifiable, separate and distinct functions of  $y_1$  and  $y_2$ , indicating that the two variables are independent), or by integrating out  $y_2$  in Eq (6.95), one easily obtains the required marginal pdf for  $Y_1$  as:

$$f_1(y_1) = \frac{1}{2\sqrt{\pi}} e^{-\frac{y_1^2}{4}}; -\infty < y_1 < \infty \quad (6.96)$$

In the next example we derive one more important result and illustrate the seriousness of the requirement that the Jacobian of the inverse transformation not vanish anywhere in  $V_Y$ .

**Example 6.9 RATIO OF TWO STANDARD NORMAL RANDOM VARIABLES**

Given two stochastically independent random variables  $X_1$  and  $X_2$  with pdfs:

$$f_1(x_1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x_1^2}{2}}; -\infty < x_1 < \infty \quad (6.97)$$

$$f_2(x_2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x_2^2}{2}}; -\infty < x_2 < \infty \quad (6.98)$$

determine the pdf of the random variable  $Y$  obtained from their ratio,

$$Y = X_1/X_2 \quad (6.99)$$

**Solution:**

Again, because this is an underdetermined transformation, we must first augment it with another one, say  $Y_2 = X_2$ , to give:

$$\begin{aligned} Y_1 &= \frac{X_1}{X_2} \\ Y_2 &= X_2 \end{aligned} \quad (6.100)$$

which is now square, one-to-one, and with the inverse transformation:

$$\begin{aligned} x_1 &= y_1 y_2 \\ x_2 &= y_2 \end{aligned} \quad (6.101)$$

The Jacobian,

$$J = \begin{vmatrix} y_2 & y_1 \\ 0 & 1 \end{vmatrix} = y_2 \quad (6.102)$$

vanishes at the single point  $y_2 = 0$ , however; and even though this is a point of probability measure zero, the observation is worth keeping in mind.

From Example 6.8 above, the joint pdf for  $X_1$  and  $X_2$  is given by:

$$f_{\mathbf{X}}(x_1, x_2) = \frac{1}{2\pi} e^{-\left(\frac{x_1^2+x_2^2}{2}\right)}; -\infty < x_1 < \infty; -\infty < x_2 < \infty \quad (6.103)$$

from where we now obtain the joint pdf for  $Y_1$  and  $Y_2$  as:

$$f_{\mathbf{Y}}(y_1, y_2) = \frac{1}{2\pi} |y_2| e^{-\left(\frac{y_1^2 y_2^2 + y_2^2}{2}\right)}; \quad -\infty < y_1 < \infty; \quad (6.104)$$

$$-\infty < y_2 < 0; 0 < y_2 < \infty$$

The careful reader will notice two things: (i) the expression for  $f_{\mathbf{Y}}$  involves not just  $y_2$ , but its absolute value  $|y_2|$ ; and (ii) that we have excluded the troublesome point  $y_2 = 0$  from the space  $V_Y$ . These two points are related: to the left of the point  $y_2 = 0$ ,  $|y_2| = -y_2$ ; to the right,  $|y_2| = y_2$ , so that these two regions *must* be treated differently in evaluating the integral.

To obtain the marginal pdf for  $y_1$  we now integrate out  $y_2$  in Eq (6.104) over the appropriate region in  $V_Y$  as follows:

$$f_1(y_1) = \frac{1}{2\pi} \left[ \int_{-\infty}^0 -y_2 e^{-\frac{(y_1^2+1)y_2^2}{2}} dy_2 + \int_0^\infty y_2 e^{-\frac{(y_1^2+1)y_2^2}{2}} dy_2 \right] \quad (6.105)$$

which simplifies to:

$$f_1(y_1) = \frac{1}{\pi} \left[ \frac{1}{(1+y_1^2)} \right]; -\infty < y_1 < \infty \quad (6.106)$$

as the required pdf. It is important to note that in carrying out the integration implied in (6.105), the nature of the absolute value function,  $|y_2|$ , naturally forced us to exclude the point  $y_2 = 0$  because it made it impossible for us to carry out the integration from  $-\infty$  to  $\infty$  under a single integral. (Had the integral involved not  $|y_2|$ , but  $y_2$ , as an instructive exercise, the reader should try to evaluate the resulting integral from  $-\infty$  to  $\infty$ . See Exercise 6.9.)

#### 6.4.3 Non-Monotone Transformations

In general, when the multivariate transformation  $\mathbf{y} = \Phi(\mathbf{x})$  may be non-monotone but has a countable number of roots  $k$ , when written as the matrix version of Eq (6.33), i.e.,

$$\mathbf{x}_i = \Phi_i^{-1}(\mathbf{y}) = \Psi_i(\mathbf{y}); i = 1, 2, 3, \dots, k \quad (6.107)$$

if each inverse transformation  $\Psi_i$  is square, with a non-zero Jacobian  $J_i$ , then it can be shown that:

$$f_{\mathbf{Y}}(\mathbf{y}) = \sum_{i=1}^k f_{\mathbf{X}}[\Psi_i(\mathbf{y})] |J_i| \quad (6.108)$$

which is a multivariate extension of Eq (6.35).

#### 6.5 Summary and Conclusions

We have focussed attention in this chapter on the single problem of determining the pdf,  $f_Y(y)$ , of a random variable  $Y$  that has been defined as a function of another random variable,  $X$ , whose pdf  $f_X(x)$  is known. As is common with problems of such general construct, the approach used to determine the desired pdf depends on the nature of the random variable, as well as the nature of the problem itself—in this particular case, the problem

being generally more straightforward to solve for discrete random variables than for continuous ones. When the transformation involves random variable sums, it is much easier to employ the method of characteristic functions, regardless of whether the random variables involved are discrete or continuous. But beyond the special care that must be taken for continuous non-monotone transformations, the underlying principle is the same for all cases and is fairly straightforward.

The primary importance of this chapter lies in the fact that it provides one of the tools (and much of the foundational results) employed routinely in deriving probability models for some of the more complex random phenomena. We will therefore rely on much of this chapter's material in subsequent chapters, especially in Chapters 8 and 9 where we derive models for a wide variety of specific randomly varying phenomena. As such, the reader is encouraged to tackle a good number of the exercises and problems found at the end of this chapter; solving these problems will make the upcoming discussions much easier to grasp at a fundamental level.

Here are some of the main points of the chapter again.

- Given a random variable  $X$  with pdf  $f_X(x)$ , and the random variable transformation,  $Y = \phi(X)$ , the corresponding pdf  $f_Y(y)$  for the random variable  $Y$  is obtained directly from the inverse transformation,  $X = \psi(Y)$  for the discrete random variable; for continuous random variables, the Jacobian of the inverse transformation,  $J = \frac{d}{dy}[\psi(Y)]$ , is required in addition.
- When the transformation  $\phi(X)$  involves sums, it is more convenient to employ the characteristic function of  $X$  to determine  $f_Y(y)$ .
- When the transformation  $\phi(X)$  is non-monotone,  $f_Y(y)$  will consist of a sum of  $k$  components, where  $k$  is the total number of roots of the inverse transformation.
- When multivariate transformations are represented in matrix form, the required results are matrix versions of the results obtained for single variable transformations.

## REVIEW QUESTIONS

1. State, in mathematical terms, the problem of primary interest in this chapter.
2. What are the results of this chapter useful for?
3. In single variable transformations, where  $Y = \phi(X)$  is given along with  $f_X(x)$ , and  $f_Y(y)$  is to be determined, what is the difference between the discrete case of this problem and the continuous counterpart?

4. What is the Jacobian of the single variable inverse transformation?
5. In determining  $f_Y(y)$  given  $f_X(x)$  and the transformation  $Y = \phi(X)$ , how does one handle the case where the transformation is not strictly monotone?
6. Which two approaches were presented in this chapter for finding pdfs of random variable sums? Which of the two is more convenient?
7. What is meant by the “convolution of two functions,  $f_1(x_1)$  and  $f_2(x_2)$ ?”
8. Upon what property of characteristic functions is the “characteristic function approach” to the determination of the pdf of random variable sums based?
9. What is the Jacobian of a multivariate inverse transformation?
10. How are non-square transformations handled?

## EXERCISES

**6.1** The pdf of a random variable  $X$  is given as:

$$f(x) = p(1-p)^{x-1}; x = 1, 2, 3, \dots, \quad (6.109)$$

(i) Obtain the pdf for the random variable  $Y$  defined as

$$Y = \frac{1}{X} \quad (6.110)$$

(ii) Given that  $E(X) = 1/p$ , obtain  $E(Y)$  and compare it to  $E(X)$ .

**6.2** Given the pdf shown in Eq (6.18) for the transformed variable,  $Y$ , i.e.,

$$f_Y(y) = \frac{e^{-(2^y)}}{(\log_2 y)!}; y = 1, 2, 4, 8, \dots$$

show that  $E(Y) = e^\lambda$  and hence confirm Eq (6.21).

**6.3** Consider the random variable,  $X$ , with the following pdf:

$$f_X(x) = \begin{cases} \frac{1}{\beta} e^{-x/\beta}; & 0 < x < \infty \\ 0; & \text{elsewhere} \end{cases} \quad (6.111)$$

Determine the pdf for the random variable  $Y$  obtained via the transformation

$$Y = \frac{1}{\beta} e^{-X/\beta} \quad (6.112)$$

Compare this result to the one obtained in Example 6.2 in the text.

**6.4** Given a random variable,  $X$ , with the following pdf:

$$f_X(x) = \begin{cases} \frac{1}{2}(x+1); & -1 < x < 1 \\ 0; & \text{elsewhere} \end{cases} \quad (6.113)$$

- (i) Determine the pdf for the random variable  $Y$  obtained via the transformation

$$Y = X^2 \quad (6.114)$$

- (ii) Determine  $E(X)$  and  $E(Y)$ .

**6.5** Given the pdf for two stochastically independent random variables  $X_1$  and  $X_2$  as

$$f(x_i) = \frac{e^{-\lambda_i} \lambda_i^{x_i}}{x_i!}; x_i = 0, 1, 2, \dots \quad (6.115)$$

for  $i = 1, 2$ , and given the corresponding characteristic function as:

$$\varphi_{X_i}(t) = e^{[\lambda_i(e^{jt}-1)]}, \quad (6.116)$$

- (i) Obtain the pdf  $f_Y(y)$  of the random variable  $Y$  defined as the sum of these two random variables, i.e.,

$$Y = X_1 + X_2$$

- (ii) Extend the result to a sum of  $n$  such random variables, i.e.,

$$Y = X_1 + X_2 + \dots + X_n$$

with each distribution given in Eq (6.115). Hence, establish that the random variable  $X$  also possesses the “reproductive property” illustrated in Example 6.6 in the text.

- (iii) Obtain the pdf  $f_Z(z)$  of the random variable  $Z$  defined as the average of  $n$  such random variables, i.e.,

$$Z = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$$

**6.6** In Example 6.3 in the text, it was established that if the random variable  $X$  has the following pdf:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}; -\infty < x < \infty \quad (6.117)$$

then the pdf for the random variable  $Y = X^2$  is:

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-y/2} y^{-1/2}; 0 < y < \infty \quad (6.118)$$

Given that the characteristic function of this random variable  $Y$  is:

$$\varphi_Y(t) = \frac{1}{(1 - j2t)^{1/2}} \quad (6.119)$$

by re-writing  $\sqrt{2}$  as  $2^{1/2}$ , and  $\sqrt{\pi}$  as  $\Gamma(1/2)$  (or otherwise), obtain the pdf  $f_Z(z)$  of the random variable defined as:

$$Z = X_1^2 + X_2^2 + \dots + X_r^2 \quad (6.120)$$

where the random variables,  $X_i$ , are all mutually stochastically independent, and each has the distribution shown in Eq (6.117).

**6.7** Revisit Example 6.8 in the text, but this time, instead of Eq (6.91), use the following alternative “squaring” transformation,

$$Y_2 = X_2 \quad (6.121)$$

You should obtain the same result.

**6.8** Revisit Example 6.9 in the text, but this time, instead of Eq (6.100), use the following alternative “squaring” transformation,

$$Y_2 = X_1 \quad (6.122)$$

Which augmenting squaring transformation leads to an easier problem—this one, or the one in Eq (6.100) used in Example 6.8?

**6.9** Revisit Eq (6.104), this time, replace  $|y_2|$  with  $y_2$ , and integrate the resulting joint pdf  $f_{Y_1 Y_2}(y_1, y_2)$  with respect to  $y_2$  over the entire range  $-\infty < y_2 < \infty$ . Compare your result with Eq (6.106) and comment on the importance of making sure to use the *absolute value* of the Jacobian of the inverse transformation in deriving pdfs of continuous transformed variables.

## APPLICATION PROBLEMS

**6.10** In a commercial process for manufacturing the extruded polymer film Mylar®, each roll of the product is characterized in terms of its “gage,” the film thickness,  $X$ . For a series of rolls that meet the desired mean thickness target of  $350 \mu\text{m}$ , the thickness of a section of film sampled randomly from a particular roll has the pdf

$$f(x) = \frac{1}{\sigma_i \sqrt{2\pi}} \exp \left\{ \frac{-(x - 350)^2}{2\sigma_i^2} \right\} \quad (6.123)$$

where  $\sigma_i^2$  is the variance associated with the average thickness for each roll,  $i$ . In reality, the product property that is of importance to the end-user is not so much the film thickness, or even the average film thickness, but a roll-to-roll consistency, quantified in terms of a “relative thickness variability” measure defined as

$$Y = \left( \frac{X - 350}{\sigma_i} \right)^2 \quad (6.124)$$

Obtain the pdf  $f_Y(y)$  that is used to characterize the roll-to-roll variability observed in this product quality variable.

**6.11** Consider an experimental, electronically controlled, mechanical tennis ball launcher designed to be used to train tennis players. One such machine is positioned at a fixed launch point,  $L$ , located a distance of  $1 \text{ m}$  from a wall as shown in Fig 6.2. The launch mechanism is programmed to launch the ball in an essentially straight line, at an angle  $\theta$  that varies randomly according to the pdf:

$$f(\theta) = \begin{cases} c; & -\frac{\pi}{2} < \theta < \frac{\pi}{2} \\ 0; & \text{elsewhere} \end{cases} \quad (6.125)$$

where  $c$  is a constant. The point of impact on the wall, at a distance  $y$  from the

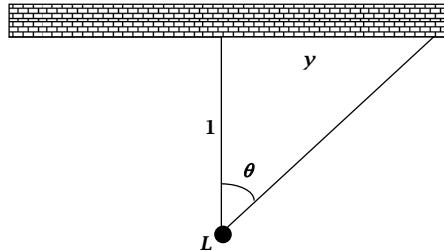


FIGURE 6.2: Schematic diagram of the tennis ball launcher of Problem 6.11

center, will therefore be a random variable whose specific value depends on  $\theta$ . First show that  $c = \pi$ , and then obtain  $f_Y(y)$ .

**6.12** The distribution of residence times in a single continuous stirred tank reactor (CSTR), whose volume is  $V$  liters and through which reactants flow at rate  $F$  liters/hr, was established in Chapter 2 as the pdf:

$$f(x) = \frac{1}{\tau} e^{-x/\tau}; 0 < x < \infty \quad (6.126)$$

where  $\tau = V/F$ .

(i) Find the pdf  $f_Y(y)$  of the residence time,  $Y$ , in a reactor that is 5 times as large, given that in this case,

$$Y = 5X \quad (6.127)$$

(ii) Find the pdf  $f_Z(z)$  of the residence time,  $Z$ , in an ensemble of 5 reactors in series, given that:

$$Z = X_1 + X_2 + \cdots + X_5 \quad (6.128)$$

where each reactor's pdf is as given in Eq (6.126), with parameter,  $\tau_i; i = 1, 2, \dots, 5$ . (Hint: Use results of Examples 6.5 and 6.6).

(iii) Show that even if  $\tau_1 = \tau_2 = \cdots = \tau_5 = \tau$  for the ensemble of 5 reactors in series,  $f_Z(z)$  will still *not* be the same as  $f_Y(y)$ .

**6.13** The total number of flaws (dents, scratches, paint blisters, etc) found on the various sets of doors installed on brand new minivans in an assembly plant is a random variable with the pdf:

$$f(x) = \frac{e^{-\lambda} \lambda^x}{x!}; x = 0, 1, 2, \dots \quad (6.129)$$

The value of the pdf parameter,  $\lambda$ , depends on the door in question as follows:  $\lambda = 0.5$  for the driver and front passenger doors;  $\lambda = 0.75$  for the two bigger mid-section passenger doors, and  $\lambda = 1.0$  for the fifth, rear trunk/tailgate door. If the total number of flaws per completely assembled minivan is  $Y$ , obtain the pdf  $f_Y(y)$  and from it, compute the probability of assembling a minivan with more than a total number of 2 flaws on all its doors.

**6.14** Let the fluorescence signals obtained from a test spot and the reference spot

on a microarray be represented as random variables  $X_1$  and  $X_2$  respectively. Within reason, these variables can be assumed to be independent, with the following pdfs:

$$f_1(x_1) = \frac{1}{\Gamma(\alpha)} x_1^{\alpha-1} e^{-x_1}; 0 < x_1 < \infty \quad (6.130)$$

$$f_2(x_2) = \frac{1}{\Gamma(\beta)} x_2^{\beta-1} e^{-x_2}; 0 < x_2 < \infty \quad (6.131)$$

It is customary to analyze such microarray data in terms of the “fold change” ratio,

$$Y = \frac{X_1}{X_2} \quad (6.132)$$

indicative of the “fold increase” (or decrease) in the signal intensity between test and reference conditions. Show that the pdf of  $Y$  is given by:

$$f(y) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{y^{\alpha-1}}{(1+y)^{\alpha+\beta}}; y > 0; \alpha > 0; \beta > 0 \quad (6.133)$$

**6.15** The following expression is used to calibrate a thermocouple whose natural output is  $V$  volts;  $X$  is the corresponding temperature, in degrees Celsius.

$$X = 0.4V + 100 \quad (6.134)$$

in a range from 50 to 500 volts and 100 to 250 °C. If the voltage output is subject to random variability around the true value  $\mu_V$ , such that

$$f(v) = \frac{1}{\sigma_V \sqrt{2\pi}} \exp \left\{ \frac{-(v - \mu_V)^2}{2\sigma_V^2} \right\} \quad (6.135)$$

where the mean (i.e., expected) value for Voltage,  $E(V) = \mu_V$  and the variance,  $Var(V) = \sigma_V^2$ , (i) Show that:

$$E(X) = 0.4\mu_V + 100 \quad (6.136)$$

$$Var(X) = 0.16\sigma_V^2 \quad (6.137)$$

(ii) In terms of  $E(X) = \mu_X$  and  $Var(x) = \sigma_X^2$ , obtain an expression for the pdf  $f_X(x)$  representing the variability propagated to the temperature values.

**6.16** “Propagation-of-errors” studies are concerned with determining how the errors from one variable are transmitted to another when the two variables are related according to a known expression. When the relationships are linear, it is often possible to obtain complete probability distribution functions for the dependent variable given the pdf for the independent variable (see Problem 6.15). When the relationships are nonlinear, closed form expressions are not always possible; in terms of general results, the best one can hope for are approximate expressions for the expected value and variance of the dependent variable, typically in a local region, upon linearizing the nonlinear expression. The following is an application of these principles.

One of the best known laws of bioenergetics, Kleiber’s law, states that the “Resting Energy Expenditure” of an animal,  $Q_0$ , (essentially the animal’s metabolic rate,

in kcal/day), is proportional to  $M^{3/4}$ , where  $M$  is the animal's mass (in kg). Specifically for mature homeotherms, the expression is:

$$Q_0 = 70M^{3/4} \quad (6.138)$$

Consider a particular population of homeotherms for which the variability in mass is characterized by the random variable  $M$  with the distribution:

$$f(m) = \frac{1}{\sigma_M \sqrt{2\pi}} \exp \left\{ \frac{-(m - M^*)^2}{2\sigma_M^2} \right\} \quad (6.139)$$

with a mean value,  $M^*$ , and variance  $\sigma_M^2$ . The pdf representing the corresponding variation in  $Q_0$  can be obtained using the usual transformation techniques, but the result does not have a convenient, recognizable, closed form. However, it is possible to obtain approximate values for  $E(Q_0)$  and  $Var(Q_0)$  in the neighborhood around the mean mass,  $M^*$ , and the corresponding metabolic rate,  $Q_0^*$ .

Given that a first-order (linear) Taylor series approximation of the expression in Eq (6.138) is defined as:

$$Q_0 \approx Q_0^* + \left. \frac{\partial Q_0}{\partial M} \right|_{M=M^*} (M - M^*) \quad (6.140)$$

first obtain the approximate linearized expression for  $Q_0$  when  $M^* = 75$  kg, and then determine  $E(Q_0)$  and  $Var(Q_0)$  for a population with  $\sigma_M = 12.5$  kg under these conditions.



# Chapter 7

## *Application Case Studies I: Probability*

|       |   |     |
|-------|---|-----|
| 7.1   | Introduction .....  | 198 |
| 7.2   | Mendel and Heredity .....                                 | 199 |
| 7.2.1 | Background and Problem Definition .....                   | 199 |
| 7.2.2 | Single Trait Experiments and Results .....                | 200 |
| 7.2.3 | Single trait analysis .....                               | 201 |
|       | The First Generation Traits .....                         | 203 |
|       | Probability and The Second Generation Traits .....        | 204 |
| 7.2.4 | Multiple Traits and Independence .....                    | 205 |
|       | Pairwise Experiments .....                                | 205 |
| 7.2.5 | Subsequent Experiments and Conclusions .....              | 208 |
| 7.3   | World War II Warship Tactical Response Under Attack ..... | 209 |
| 7.3.1 | Background and Problem Definition .....                   | 209 |
| 7.3.2 | Approach and Results .....                                | 209 |
| 7.3.3 | Final Comments .....                                      | 212 |
| 7.4   | Summary and Conclusions .....                             | 212 |

*But to us, probability is the very guide of life.*

Bishop Butler (1692–1752)

To many scientists and engineers, a first encounter with the theory of probability in its modern axiomatic form often leaves the impression of a subject matter so abstract and esoteric in nature as to be entirely suited to nothing but the most contrived applications. Nothing could be further from the truth. In reality, the application of probability theory features prominently in many modern fields of study: from finance, economics, sociology and psychology to various branches of physics, chemistry, biology and engineering, providing a perfect illustration of the aphorism that “there is nothing so practical as a good theory.”

This chapter showcases the applicability of probability theory through two specific case studies involving real-world problems whose practical importance can hardly be overstated. The first, Mendel’s deduction of the laws of heredity—the basis for the modern science of genetics—shows how Mendel employed probability (and the concept of “stochastic independence”) to establish the principles underlying a phenomenon which, until then, was considered essentially unpredictable and hence not susceptible to systematic analysis.

The second is from a now-declassified US Navy study during World War II and involves decision-making in the face of uncertainty, using past data. It

illustrates the application of frequency-of-occurrence information, viewed as approximate total and conditional probabilities, to solve an important tactical military problem.

---

## 7.1 Introduction

The elegant, well-established and fruitful tree we now see as modern probability theory has roots that reach back to 16<sup>th</sup> and 17<sup>th</sup> century gamblers and the very real—and very practical—need for reliable solutions to numerous gambling problems. Referring to these gambling problems by the somewhat less morally questionable term “problems on games of chance,” some of the most famous and most gifted mathematicians of the day devoted considerable energy first to solving specific problems (most notably the Italian mathematician, Cardano, in the 16<sup>th</sup> century), and later to developing the foundational basis for systematic mathematical analysis (most notably the Dutch scientist, Huygens, and the French mathematicians, Pascal and Fermat, in the 17<sup>th</sup> century). However, despite subsequent major contributions in the 18<sup>th</sup> century from the likes of Jakob Bernoulli (1654-1705) and Abraham de Moivre (1667-1754), it was not until the 19<sup>th</sup> century, with the publication in 1812 of Laplace’s book, *Théorie Analytique des Probabilités*, that probability theory moved beyond the mathematical analysis of games of chance to become recognized as an important branch of mathematics in its own right—one with broader applications to other scientific and practical problems such as statistical mechanics and characterization of experimental error.

The final step in the ascent of probability theory was taken in the 20<sup>th</sup> century with the development of the axiomatic approach. First expounded in Kolmogorov’s celebrated 1933 monograph (the English translation, *Foundations of Probability Theory*, was published in 1950), this approach, once and for all, provided a rigorous and mathematically precise definition of probability that sufficiently generalized the theory and formalized its applicability to a wide variety of random phenomena. Paradoxically, that probability theory in its current modern form enjoys applications in such diverse areas as actuarial science, economics, finance; genetics, medicine, psychology; engineering, manufacturing, and strategic military decisions, is attributable to Kolmogorov’s rigorous theoretical and precise formalism. Thus, even though it would be considered overly hyperbolic today (too much embroidery), placed in its proper historic context, the following statement in Laplace’s book is essentially true:

*“It is remarkable that a science which began with the consideration of games of chance should have become the most important object of human knowledge.”*

The two example case studies presented here illustrate just how important probability theory and its application have become since the time of Laplace.

## 7.2 Mendel and Heredity

Heredity, how traits are transmitted from parent to offspring, has always fascinated — and puzzled — mankind. This phenomenon, central to the propagation of life itself, with serious implications for the health, viability and survival of living organisms, remained mysterious and poorly understood until the ground-breaking work by Gregor Mendel (1822-1884). Mendel, an Augustinian monk, arrived at his amazing conclusions by studying variations in pea plants, using the garden of his monastery as his laboratory. As stated in an English translation of the original paper published in 1866<sup>1</sup> “... The object of the experiment was to observe these variations in the case of each pair of differentiating characters, and to deduce the law according to which they appear in successive generations.” The experiments involved the careful cultivation and testing of nearly 30,000 plants, lasting almost 8 years, from 1856 to 1863. The experimental results, and their subsequent probabilistic analysis paved the way for the modern science of genetics, but it was not recognized as such right away. The work, and its monumental import, languished in obscurity until the early 20<sup>th</sup> century when it was rediscovered and finally accorded its well-deserved recognition.

What follows is an abbreviated discussion of the essential elements of Mendel’s work and the probabilistic reasoning that led to the elucidation of the mechanisms behind heredity and genetics.

### 7.2.1 Background and Problem Definition

*“The value and utility of any experiment are determined by the fitness of the material to the purpose for which it is used, and thus in the case before us it cannot be immaterial what plants are subjected to experiment and in what manner such experiment is conducted.”*

So wrote Mendel in motivating his choice of pea plants as the subject of his now-famous set of experiments. The two primary factors that made pea plants an attractive choice are:

1. Relatively fast rates of reproduction; and

<sup>1</sup>Mendel, Gregor, 1866. Versuche über Pflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brünn*, Bd. IV für das Jahr 1865, Abhandlungen, 347; first translated into English by William Bateson in 1901 as “Experiments in Plant Hybridization,”: see <http://www.netSPACE.org/MendelWeb/>.

2. Availability of many varieties, each producing definite and easy to characterize traits;

Before enumerating the specific traits that Mendel studied, it is important to note in hindsight, that the choice of peas was remarkably fortuitous because the genetic structure of peas is now known to be relatively simple. A more complex genetic structure could have further obscured the fundamental principles with additional distracting details; and the deductive analysis required to derive general laws governing heredity and genetics from this specific set of results would have been far more difficult.

By tracking the following seven specific trait characteristics (with the variations manifested in each trait indicated in square brackets),

1. Seed Shape; [**Round**/ *Wrinkled*]
2. Seed Albumen Color; [**Yellow**/ *Green*]
3. Seed Coat (same as Flower); Color [**Reddish**/ *White*]
4. Pod Form (or Texture); [**Inflated**/ *Constricted*]
5. Unripe Pod (or stalks) Color; [**Green**/ *Yellow*]
6. Flower Position (on the stem); [**Axial**/ *Terminal*]
7. Stem length; [**Tall**/ *Dwarfed*]

Mendel sought to answer the following specific questions:

1. How are these seven traits transmitted from parents to offsprings generation after generation?
2. Are there discernible patterns and can they be generalized?

Our discussion here is limited to two out of the many sets of experiments in the original study:

1. *Single trait experiments*, in which individual traits and how they are transmitted from parent to offspring in subsequent generations are tracked one-at-a-time;
2. *Multiple trait experiments*, in which several traits and their transmission are tracked simultaneously, specifically focusing on pairwise experiments involving two traits.

### 7.2.2 Single Trait Experiments and Results

In this set of experiments Mendel focussed on individual traits (such as seed shape) one-at-a-time, and tracked how they are transmitted from one generation to the next. In all, seven different sets of experiments were performed, each set devoted to a single trait, but for the sake of the current presentation, we will use seed shape as a representative example.

The experimental procedure and results are summarized as follows:

1. *Initialization: Generation of “pure” traits parents ( $P$ )*. This was done by fertilizing pollen from round seed shaped plants for many generations until they “stabilized” and produced only round seed offsprings consistently. The same procedure was repeated for wrinkled seeds. (The other 6 traits received the same treatment for each pair of associated trait variations)
2. *First generation hybrids ( $f_1$ ): Cross-fertilization of “pure” traits*. Pure parent round seed plants were cross-fertilized with wrinkled ones to produce first generation hybrids. Results: from parents with separate and distinct “pure” traits, every single seed in the first generation of hybrids was round, not a single wrinkled! (Identical corresponding results were obtained for the other 6 traits, with one variant manifesting preferentially over the complementary pair).
3. *Second generation hybrids ( $f_2$ ) from the first generation*. The first generation plants were cross-fertilized among themselves. Result: approximately 3/4 of the seeds were round, with 1/4 wrinkled. (Identical corresponding results were obtained for the other 6 traits: the variant that was exclusively manifested in the first generation hybrids retained its preferential status, and did so in the same approximate 3:1 ratio.)

The entire collection of results from all seven sets of experiments are summarized in Table 7.1.

### 7.2.3 Single trait analysis

To make sense of these surprising and somewhat counterintuitive results required convincing answers to the following fundamental questions raised by the data:

1. *First generation visible trait uniformity*: How does the cross-fertilization of round seed plants with wrinkled seed ones produce a first generation of hybrid seeds that are *all* round? Or, in general, how does the cross-fertilization of two very different, pure, and distinct traits produce first generation hybrid offsprings that all—without exception—preferentially display only one of these parental traits?
2. *Second generation visible trait variety*: How does the cross-fertilization

**TABLE 7.1:** Summary of Mendel's single trait experiment results

| Characteristics                     | 1 <sup>st</sup> Generation |          |           | 2 <sup>nd</sup> Generation |                          |        | D : r Ratio |
|-------------------------------------|----------------------------|----------|-----------|----------------------------|--------------------------|--------|-------------|
|                                     | Total                      | Dominant | Recessive | Proportion Dominant (D)    | Proportion Recessive (r) |        |             |
| Seed Shape<br>(Round/Wrinkled)      | Round                      | 7,324    | 5,474     | 1,850                      | 0.747                    | 0.253  | 2.96:1      |
| Seed Alb Color<br>(Yellow/Green)    | Yellow                     | 8,023    | 6,022     | 2,001                      | 0.751                    | 0.249  | 3.01:1      |
| Seed Coat/Flower<br>(Reddish/White) | Reddish                    | 929      | 705       | 224                        | 0.759                    | 0.241  | 3.15:1      |
| Pod Form<br>(Inflated/Constricted)  | Inflated                   | 1,181    | 882       | 299                        | 0.747                    | 0.253  | 2.95:1      |
| Unripe Pod Color<br>(Green/Yellow)  | Green                      | 580      | 428       | 152                        | 0.738                    | 0.262  | 2.82:1      |
| Flower Position<br>(Axial/Terminal) | Axial                      | 858      | 651       | 207                        | 0.759                    | 0.241  | 3.14:1      |
| Stem Length<br>(Tall/Dwarfed)       | Tall                       | 1,064    | 787       | 277                        | 0.740                    | 0.260  | 2.84:1      |
| Totals                              | 19,959                     | 14,949   | 5010      | 0.749                      | 0.251                    | 2.98:1 |             |

of first generation plants (with only one trait uniformly on display) produce second generation plants displaying a variety entirely absent in the homogenous first generation? Or, alternatively, how did the missing trait in the first generation reappear in the second?

3. *Second generation visible trait composition:* What law governs the apparently constant numerical ratio with which the original parental traits appear in the second generation? What is the “true” theoretical value of this numerical ratio?

To answers these questions and elucidate the principles governing single trait selection, Mendel developed the following concepts and demonstrated that they were consistent with his experimental data:

#### 1. The concept of Hereditary Factors:

- The inheritance of each trait is determined by “units” or “factors” (now called genes); these “factors” do not amalgamate, but are passed on to offsprings intact and unchanged;
- An individual has two sets of such units or factors, inheriting one set from each parent; thus each parent transmits only half of its hereditary factors to each offspring;
- Which of the two parental factors is inherited by an offspring is purely a matter of chance.

#### 2. The concept of Dominance/Recessiveness:

- In heredity, one trait is always dominant over the other, this *other* trait being the recessive one;
- To “show up,” a dominant trait needs only one trait factor from the parent; the recessive needs two;
- A trait may not show up in an individual but its factor can still be transmitted to the next generation.

Mendel’s postulate was that if these concepts are true, then one must obtain the observed results; conversely one will obtain these results *only* if these concepts are valid.

### The First Generation Traits

To see how these concepts help resolve the first problem, consider first the specific case of the seed shape: Let the “factors” possessed by the pure round shaped parent be represented as **RR**, (each **R** representing one round trait “factor”); similarly, let the factors of the pure wrinkled shaped parent be represented as **ww**. In cross-fertilizing the round-seed plants with the wrinkled-seed ones, each first generation hybrid will have factors that are either **Rw** or **wR**. And now, if the “round” trait is dominant over the wrinkled trait,

then observe that the entire first generation will be all round, precisely as in Mendel's experiment.

In general, when a pure dominant trait with factors **DD** is cross-fertilized with a pure recessive trait with factors **rr**, the first generation hybrid will have factors **Dr** or **rD** each one displaying uniformly the dominant trait, but carrying the recessive trait. The concepts of hereditary factors (genes) and of dominance thus enabled Mendel to resolve the problem of the uniform display of traits in the first generation; just as important, they also provided the foundation for elucidating the principles governing trait selection in the second and subsequent generations. This latter exercise is what would require probability theory.

### Probability and The Second Generation Traits

The key to the second generation trait manifestation is a recognition that while each seed of the first generation plants "looks" like the dominant round-seed type in the parental generation, there are some fundamental, but invisible, differences: the parental generation has pure trait factors **RR** and **ww**; the first generation has two distinct trait factors: **Rw** (or **wR**), one visible (phenotype) because it is dominant, the other *not* visible but inherited nonetheless (genotype). The hereditary but otherwise invisible trait is the key.

To analyze the composition of the second generation, the following is a modernization of the probabilistic arguments Mendel used. First note that the collection of all possible outcomes when cross-fertilizing two plants each with a trait factor set **Rw** is given by:

$$\Omega = \{\mathbf{RR}, \mathbf{Rw}, \mathbf{wR}, \mathbf{ww}\} \quad (7.1)$$

From here, according to the theory of hereditary factors and dominance, it should now be clear that there will be a mixture of round seeds as well as wrinkled seeds. But because it is purely a matter of chance which factor is passed on from the first generation to the next, this set is rightly considered the "sample space" of the experiment. (In fact, the phenomenon in question is precisely akin to the idealized experiment in which a coin is tossed twice and the number of heads and tails are recorded, for example with *H* as **R**, and *T* as **w**.)

To determine the ratio in which these phenotypic traits will be displayed, let the random variable of interest (in this case the manifested phenotypic trait) be defined as follows:

$$X = \begin{cases} 0, & \text{Wrinkled} \\ 1, & \text{Round} \end{cases} \quad (7.2)$$

in which case,

$$V_X = \{0, 1\}; \quad (7.3)$$

If the theory of dominance is valid, and if there is an equiprobable chance

for each trait combination, then from  $V_x$  and its pre-image in  $\Omega$ , Eq. (7.1), the probability distribution function of the phenotypic manifestation random variable,  $X$ , is given by

$$P(X = 0) = 1/4 \quad (7.4)$$

$$P(X = 1) = 3/4 \quad (7.5)$$

The second generation composition will therefore be a 3:1 ratio of round to wrinkled seeds. (The same arguments presented above apply to all the other single traits.)

Placed side-by-side with the experimental results shown in Table 7.1, these probabilistic arguments are now seen to confirm all the postulated concepts and theories. The fact that the dominant-to-recessive trait ratios observed experimentally did not come out to be precisely 3:1 in all the traits is of course a consequence of the random fluctuations intrinsic to all random phenomena. Note also how the ratios determined from larger experimental samples ( $\sim 7,000$  and  $\sim 8,000$  respectively for shape and albumen color) are closer to the theoretical value than the ratios obtained from much smaller samples ( $\sim 600$  and  $\sim 800$  respectively for pod color and flower position). These facts illustrate the fundamental difference between empirical frequencies and theoretical probabilities: the former will not always match the latter exactly, but the difference will dwindle to nothingness as the sample size increases, with the two coinciding in the limit of infinite sample size. The observed results are akin to the idealized experiment of tossing a fair coin twice, and determining the number of time one obtains at least a “Head”. Theoretically, this event should occur 3/4 of the time, but there will be fluctuations.

#### 7.2.4 Multiple Traits and Independence

The discussion thus far has been concerned with single traits and the principles governing their hereditary transmission. Mendel's next task was to determine whether these principles applied equally to trait pairs, and then in general “when several diverse characters are united in the hybrid by crossing.” The key question to be answered—“does the transmission of one trait interfere with another, or are they wholly independent”—required a series of carefully designed experiments on a large number of plants.

#### Pairwise Experiments

The first category of multiple trait experiments involved cross-fertilization of plants in which the differentiating characteristics were considered in pairs. For the purpose of illustration, we will consider here only the very first in this series, in which the parental plants differed in seed *shape* and seed *albumen color*. Specifically, the seed plants were round and yellow (**R,Y**), while the pollen plants were wrinkled and green (*w,g*). (To eliminate any possible systematic “pollen” or “seed” effect, Mendel also performed a companion series of

experiments in which the roles of seed and pollen were reversed.) The specific question the experiments were designed to answer is this: *will the transmission of the shape trait interfere with color or will they be independent?*

As in the single trait experiments, the first generation of hybrids were obtained by cross-fertilizing the pure round-and-yellow seed plants with pure wrinkled-and-green ones; the second generation plants were obtained by cross-fertilizing first generation plants, and so on, with each succeeding generation similarly obtained from the immediately preceding one.

**First generation results:** The first generation of fertilized seeds ( $f_1$ ) were *all* round and yellow like the seed parents. These results are definitely reminiscent of the single trait experiments and appeared to confirm that the principle of dominance extended to pairwise traits independently: i.e. that the round shape trait dominance over the wrinkled, *and* the yellow color trait dominance over the green held true in the pairwise experiments just as they did in the single trait experiments. Shape did not seem to interfere with color, at least in the first generation. But how about the second generation? How will the bivariate shape/color traits manifest, and how will this influence the composition of the second generation hybrids? The circumstances are clearly more complicated and require more careful analysis.

**Second generation: Postulate, Theoretical Analysis and Results:** Rather than begin with the experimental results and then wend our way through the theoretical analysis required to explain the observations, we find it rather more instructive to begin from a postulate, and consequent theoretical analysis, and proceed to compare the theoretical predictions with experimental data.

As with the single trait case, let us define the following random variables: for shape,

$$X_1 = \begin{cases} 0, & \text{Wrinkled} \\ 1, & \text{Round} \end{cases} \quad (7.6)$$

and for color,

$$X_2 = \begin{cases} 0, & \text{Green} \\ 1, & \text{Yellow} \end{cases} \quad (7.7)$$

As obtained previously, the single trait marginal pdfs for second generation hybrid plants are given by:

$$f_1(x_1) = \begin{cases} 1/4; & x_1 = 0 \\ 3/4; & x_1 = 1 \end{cases} \quad (7.8)$$

for shape, and, similarly, for color,

$$f_2(x_2) = \begin{cases} 1/4; & x_2 = 0 \\ 3/4; & x_2 = 1 \end{cases} \quad (7.9)$$

We now desire the joint pdf  $f(x_1, x_2)$ .

**TABLE 7.2:** Theoretical distribution of shape-color traits in second generation hybrids under the independence assumption

| Shape ( <i>w/R</i> )<br><i>X</i> <sub>1</sub> | Color ( <i>g/Y</i> )<br><i>X</i> <sub>2</sub> | Prob Dist.<br><i>f</i> ( <i>x</i> <sub>1</sub> , <i>x</i> <sub>2</sub> ) | Phenotype Trait |
|---|---|--|-----------------|
| 0   | 0   | 1/16   | ( <i>w,g</i> )  |
| 0   | 1   | 3/16   | ( <i>w,Y</i> )  |
| 1   | 0   | 3/16   | ( <b>R,g</b> )  |
| 1   | 1   | 9/16   | ( <b>R,Y</b> )  |

Observe that the set of possible trait combinations is as follows:

$$\Omega = \{(w,g), (w,Y), (\mathbf{R},g), (\mathbf{R},Y)\} \quad (7.10)$$

giving rise to the 2-dimensional random variable space:

$$V_X = \{(0,0), (0,1), (1,0), (1,1)\}. \quad (7.11)$$

Consider first the simplest postulate that multiple trait transmissions are independent. If this is true, then by definition of “stochastic independence” the joint pdf will be given by:

$$f(x_1, x_2) = f_1(x_1)f_2(x_2) \quad (7.12)$$

so that the theoretical distribution of the second generation hybrids will be as shown in Table 7.2, predicting that the observed traits will be in the proportion of 9:3:3:1 with the round-and-yellow variety being the most abundant, the wrinkled-and-green the least abundant, and the wrinkled-and-yellow and the round-and-green in the middle, appearing in equal numbers.

Mendel’s experimental results were summarized as follows:

“The fertilized seeds appeared round and yellow like those of the seed parents. The plants raised therefrom yielded seeds of four sorts, which frequently presented themselves in one pod. In all, 556 seeds were yielded by 15 plants, and of these there were:

- 315 round and yellow,
- 101 wrinkled and yellow,
- 108 round and green,
- 32 wrinkled and green.”

and a side-by-side comparison of the theoretical with the experimentally observed distribution is now shown in Table 7.3.

**TABLE 7.3:** Theoretical versus experimental results for second generation hybrid plants

| Phenotype Trait | Theoretical Distribution | Experimental Frequencies |
|-----------------|--------------------------|--------------------------|
| (R,Y)           | 0.56                     | 0.57                     |
| (w,Y)           | 0.19                     | 0.18                     |
| (R,g)           | 0.19                     | 0.19                     |
| (w,g)           | 0.06                     | 0.06                     |

Since the experimental results matched the theoretical predictions remarkably well, the conclusion is that indeed the transmission of color is independent of the transmission of the shape trait.

### 7.2.5 Subsequent Experiments and Conclusions

A subsequent series of experiments, first on other pairwise traits (yielding results similar to those shown above), and then on various simultaneous combinations of three, four and more multiple traits) led Mendel to conclude as follows:

"There is therefore no doubt that for the whole of the characters involved in the experiments the principle applies that the offspring of the hybrids in which several essentially different characters are combined exhibit the terms of a series of combinations, in which the developmental series for each pair of differentiating characters are united. It is demonstrated at the same time that the relation of each pair of different characters in hybrid union is independent of the other differences in the two original parental stocks."

Almost a century and a half after, with probability theory now a familiar fixture in the scientific landscape, and with the broad principles and consequences of genetics part of popular culture, it may be difficult for modern readers to appreciate just how truly revolutionary Mendel's experiments and his application of probability theory were. Still, it was the application of probability theory that made it possible for Mendel to predict, ahead of time, the ratio between phenotypic (visible) occurrences of dominant traits and recessive traits that will arise from a given set of parent genotype (hereditary) traits (although by today's standards all this may now appear routine and straightforward). By thus unraveling the mysteries of a phenomenon that was essentially considered unpredictable and due to chance alone plus some vague

averaging process, Mendel is rightfully recognized for laying the foundation for modern mathematical biology, influencing the likes of R. A. Fisher, and almost single-handedly changing the course of biological research permanently.

---

### 7.3 World War II Warship Tactical Response Under Attack

Unlike in the previous example where certain systematic underlying biological mechanisms were responsible for the observations, one often must deal with random phenomena for which there are no such easily discernible mechanisms behind the observations. Such is the case with the problem we are about to discuss, involving Japanese suicide bomber plane attacks during World War II. It shows how historical data sets were used to estimate empirical conditional probabilities and these probabilities subsequently used to answer a very important question with significant consequences for US Naval operations at that crucial moment during the war.

#### 7.3.1 Background and Problem Definition

During World War II, US warships attacked by Japanese kamikaze pilots had two mutually exclusive tactical options: sharp evasive maneuvers to elude the attacker and confound his aim, making a direct hit more difficult to achieve; or offensive counterattack using anti-aircraft artillery. The two options are mutually exclusive because the effectiveness of the counter attacking aircraft guns required maintaining a steady course—presenting an easier target for the incoming kamikaze pilot; sharp maneuvering warships on the other hand are entirely unable to aim and deploy their anti-aircraft guns with much effectiveness. A commitment to one option therefore immediately precludes the other.

Since neither tactic was perfectly effective in foiling kamikaze attacks, and since different types of warships—cruisers, air craft carriers, destroyers, etc—appeared to experience varying degrees of success with the different options, naval commanders needed a definitive, rational system for answering the question: *When attacked by Japanese suicide planes, what is the appropriate tactic for a US Warship, evasive maneuvers or offensive counterattack?*

**TABLE 7.4:** Attacks and hits on US WW II Naval Warships in 1943

| TACTIC                  | Large Ships (L) |           | Small Ships (S) |           | Total      |            |
|-------------------------|-----------------|-----------|-----------------|-----------|------------|------------|
|                         | Attacks         | Hits      | Attacks         | Hits      | Attacks    | Hits       |
| Evasive maneuvers       | 36              | 8         | 144             | 52        | 180        | 60         |
| Offensive Counterattack | 61              | 30        | 124             | 32        | 185        | 62         |
| <b>Total</b>            | <b>97</b>       | <b>38</b> | <b>268</b>      | <b>84</b> | <b>365</b> | <b>122</b> |

### 7.3.2 Approach and Results

The question was answered in a Naval department study commissioned in 1943 and published in 1946<sup>2</sup> although it was classified until about 1960<sup>3</sup>.

**Data:** In the summer of 1943, 365 warships that had been attacked by Kamikaze pilots provided the basis for the study and its recommendations. The data record on these ships showed warship type (Aircraft carriers, Battleships, Cruisers, Destroyers and auxiliaries), the tactic employed (evasive or offensive) and whether or not the ship was hit.

As in most cases, the raw data remains largely uninformative until appropriately reorganized. In this case, the warships were divided into two categories: “Large” (Aircraft carriers, Battleships, Cruisers) and “Small” (Destroyers and auxiliaries) and the data sorted according to the tactic employed and the corresponding number of attacks and the resulting number of hits suffered. The results are shown in Table 7.4.

**Analysis:** Assume that the data set is large enough so that frequencies can be considered as reasonable estimates of probabilities. Now consider an “experiment” consisting of selecting a warship at random; the various possible outcomes are as follows: *Ship type*:  $L$ , ship is large,  $S$  ship is small; *Naval tactic*:  $E$ , ship made an evasive maneuver,  $C$ , ship counterattacked; *Ship final status*:  $H$ , ship was hit. The problem may now be cast as that of computing probabilities of various appropriate events using the data as presented in Table 7.4, and interpreting the results accordingly.

Among the various probabilities that can be computed from this table, from the perspective of the Naval commanders, the following are the most important:

1.  $P(H)$ ,  $P(H|E)$ , and  $P(H|C)$

Respectively, the unconditional probability of any warship getting hit (i.e. the overall effectiveness of kamikaze attacks); the probability of getting hit when taking evasive measures, and when counterattacking, all regardless of size;

2.  $P(H|L)$ ;  $P(H|S)$

<sup>2</sup>P.M. Morse and G.E. Kimball, “Methods of Operations Research,” Office of the Chief of Naval Operations, Navy Department, Washington DC, 1946, Chapter 5.

<sup>3</sup>cf. R. Coughlin and D.E. Zitarelli, *The Ascent of Mathematics*, McGraw-Hill, NY, 1984, p396.

Respectively, the probability of a large ship or a small ship getting hit regardless of tactics employed;

3.  $P(H|L \cap E)$ ;  $P(H|L \cap C)$ ;

The probability of getting hit when a large ship is taking evasive maneuvers versus when counterattacking;

4.  $P(H|S \cap E)$ ;  $P(H|S \cap C)$ ;

the probability of getting hit when a small ship is taking evasive maneuvers versus when counterattacking.

These probabilities are easily computed from the table, yielding the following results:

$$P(H) = 122/365 = 0.334 \quad (7.13)$$

indicating that in general, about one in 3 attacked ships were hit, regardless of size or survival tactic employed. Similarly,

$$P(H|E) = \frac{P(H \cap E)}{P(E)} = \frac{60/365}{180/365} = 0.333 \quad (7.14)$$

(or simply directly from the rightmost column in the table). Also,

$$P(H|C) = 62/185 = 0.335 \quad (7.15)$$

The obvious conclusion: overall, there appears to be no difference between the effectiveness of evasive maneuvers as opposed to offensive counterattacks *when all the ships are considered together regardless of size*. But does size matter?

Taking size into consideration (regardless of survival tactics) the probabilities are as follows:

$$P(H|L) = 38/97 = 0.392 \quad (7.16)$$

$$P(H|S) = 84/268 = 0.313 \quad (7.17)$$

so that it appears as if small ships have a slight edge in surviving the attacks, regardless of tactics employed. But it is possible to refine these probabilities further by taking *both* size and tactics into consideration, as follows:

For large ships, we obtain

$$P(H|L \cap E) = 8/36 = 0.222 \quad (7.18)$$

$$P(H|L \cap C) = 30/61 = 0.492 \quad (7.19)$$

where we see the first clear indication of an advantage: large ships making evasive maneuvers are more than twice as effective in avoiding hits as their counterattacking counterparts.

For small ships,

$$P(H|S \cap E) = 52/144 = 0.361 \quad (7.20)$$

$$P(H|S \cap C) = 32/124 = 0.258 \quad (7.21)$$

and while the advantage is not nearly as dramatic as with large ships, it is still quite clear that small ships are more effective when counterattacking.

The final recommendation is now clear:

*When attacked by Japanese suicide planes, the appropriate tactic for small ships is to hold a steady course and counterattack; for large ships, sharp evasive maneuvers are more effective.*

### 7.3.3 Final Comments

In hindsight these conclusions from the Naval study are perfectly logical, almost obvious; but at the time, the stakes were high, time was of the essence and nothing was clear or obvious. It is gratifying to see the probabilistic analysis bring such clarity and yield results that are in perfect keeping with common sense *after the fact*.

---

## 7.4 Summary and Conclusions

This chapter, the first of a planned trilogy of case studies, (see Chapters 11 and 20 for the others) has been concerned with demonstrating the application of probability theory to two specific historical problems, each with its own significant practical implication that was probably not evident at the time the work was being done. The first, Mendel's ground-breaking work on genetics, is well-structured, and showed how a good theory can help clarify confusing experimental data. The second, the US Navy's analysis of kamikaze attack data during WW II, is less structured. It demonstrates how data, converted to empirical probabilities, can be used to make appropriate decisions. Viewed from this distance in time, and from the generally elevated heights of scientific sophistication of today, it will be all too easy to misconstrue these applications as quaint, if not trivial. But that will be a gross mistake. The significance of these applications must be evaluated within the context of the time in history when the work was done, *vis-à-vis* the tools available to the researchers at the time. The US Naval application saved lives and irreversibly affected the course of the the war in the Pacific theater. Mendel's result did not just unravel a vexing 19<sup>th</sup> century mystery; it changed the course of biological research for good, even though it was not obvious at the time. All these were made possible by the appropriate application of probability theory.



# Part III

# Distributions

—

|

—

|

—

|

—

|

---

## Part III: Distributions

*Modeling Random Variability*

---

*From a drop of water, a logician could infer  
the possibility of an Atlantic or a Niagara . . .  
So all life is a great chain, the nature of which is known  
whenever we are shown a single link of it.*

Sherlock Holmes *A Study in Scarlet*  
Sir Arthur Conan Doyle (1859–1930)

## Part III: Distributions

*Modeling Random Variability*

- **Chapter 8:** Ideal Models of Discrete Random Variables
- **Chapter 9:** Ideal Models of Continuous Random Variables
- **Chapter 10:** Information, Entropy and Probability Models
- **Chapter 11:** Application Case Studies II: In-Vitro Fertilization

# Chapter 8

## *Ideal Models of Discrete Random Variables*

|       |   |     |
|-------|---|-----|
| 8.1   | Introduction .....  | 218 |
| 8.2   | The Discrete Uniform Random Variable .....  | 219 |
| 8.2.1 | Basic Characteristics and Model .....   | 219 |
| 8.2.2 | Applications .....  | 220 |
| 8.3   | The Bernoulli Random Variable .....   | 220 |
| 8.3.1 | Basic Characteristics .....   | 221 |
| 8.3.2 | Model Development .....   | 221 |
| 8.3.3 | Important Mathematical Characteristics .....                                      | 222 |
| 8.4   | The Hypergeometric Random Variable .....  | 222 |
| 8.4.1 | Basic Characteristics .....   | 222 |
| 8.4.2 | Model Development .....   | 223 |
| 8.4.3 | Important Mathematical Characteristics .....                                      | 223 |
| 8.4.4 | Applications .....  | 224 |
| 8.5   | The Binomial Random Variable .....  | 224 |
| 8.5.1 | Basic Characteristics .....   | 225 |
| 8.5.2 | Model Development .....   | 225 |
| 8.5.3 | Important Mathematical Characteristics .....                                      | 226 |
|       | Relation to other random variables .....  | 226 |
| 8.5.4 | Applications .....  | 227 |
|       | Inference .....   | 228 |
| 8.6   | Extensions and Special Cases of the Binomial Random Variable .....                | 230 |
| 8.6.1 | Trinomial Random Variable .....   | 230 |
|       | Basic Characteristics .....   | 230 |
|       | The Model .....   | 230 |
|       | Some Important Results .....  | 231 |
| 8.6.2 | Multinomial Random Variable .....   | 231 |
| 8.6.3 | Negative Binomial Random Variable .....   | 231 |
|       | Basic Characteristics .....   | 232 |
|       | Model Development .....   | 232 |
|       | Important Mathematical Characteristics .....                                      | 233 |
| 8.6.4 | Geometric Random Variable .....   | 234 |
|       | The Model .....   | 234 |
|       | Important Mathematical Characteristics .....                                      | 234 |
|       | Applications .....  | 234 |
| 8.7   | The Poisson Random Variable .....   | 235 |
| 8.7.1 | The Limiting Form of a Binomial Random Variable .....                             | 236 |
|       | Model Development .....   | 236 |
| 8.7.2 | First Principles Derivation .....   | 237 |
|       | Basic Characteristics .....   | 237 |
|       | Model Development .....   | 237 |
| 8.7.3 | Important Mathematical Characteristics .....                                      | 239 |
| 8.7.4 | Applications .....  | 239 |
|       | Standard Poisson Phenomena .....  | 240 |
|       | Overdispersed Poisson-like Phenomena and the Negative Binomial distribution ..... | 242 |
| 8.8   | Summary and Conclusions .....   | 243 |

|                            |     |
|----------------------------|-----|
| REVIEW QUESTIONS .....     | 244 |
| EXERCISES .....            | 247 |
| APPLICATION PROBLEMS ..... | 250 |

*All these constructions and the laws connecting them  
can be arrived at by the principle of looking for  
the mathematically simplest concepts and the link between them.*

Albert Einstein (1879–1955)

Having presented the probability distribution function,  $f(x)$ , as our mathematical function of choice for representing the ensemble behavior of random phenomena, and having examined the properties and characteristics of the generic pdf extensively in the last four chapters, it now remains to present specific probability distribution functions for some actual real-world phenomena of practical importance. We do this in each case by starting with all the relevant information about the phenomenological mechanism behind the specific random variable,  $X$ ; and, in much the same way as for deterministic phenomena, we *derive* the expression for the pdf  $f(x)$  appropriate to the random phenomenon in question. The end result is several ideal models of random variability, presented as a collection of probability distribution functions, each derived directly from—and hence explicitly linked to—the underlying random phenomenological mechanisms.

This chapter and the next one are devoted to the development and analysis of such models for some important random variables that are commonly encountered in practice, beginning here with discrete random variables.

## 8.1 Introduction

As articulated briefly in the prelude chapter (Chapter 0), it is entirely possible to develop, from first-principles phenomenological considerations, appropriate theoretical characterizations of the variability inherent to random phenomena. Two primary benefits accrue from this “first-principles” approach:

1. It acquaints the reader with the mechanistic underpinnings of the random variable and the genesis of its pdf, making it less likely that the reader will inadvertently misapply the pdf to a problem to which it is unsuited. The single most insidious trap into which unsuspecting engineers and scientists often fall is that of employing a pdf inappropriately to try and solve a problem requiring a totally different pdf: for example, attempting to use the (continuous) Gaussian pdf—simply out of familiarity—inappropriately to tackle a problem involving the (discrete) phenomenon of the *number* of occurrences of safety incidents in a manufacturing site, a natural Poisson random variable.

2. It demonstrates the principles and practice of how one goes about developing such probability models, so that should it become necessary to deal with a new random phenomenon with no pre-existing “canned” model, the reader is able to fall back on first-principles to derive, with confidence, the required model.

The modeling exercise begins with a focus on discrete random variables first in this chapter, and continuous random variables next in the following chapter. In developing these models, we will draw on ideas and concepts discussed in earlier chapters about random variables, probability, probability distributions, the calculus of probability, etc., and utilize the following model development and analysis strategy:

1. Identify basic characteristics of the problem;
2. Identify elementary events and the phenomenological mechanism for combining elementary events into complex phenomena;
3. Combine components into a probability model. In each case, the resulting model will be an expression for computing  $P(X = x)$ , where  $x$  takes on discrete values  $0, 1, 2, \dots$ ;
4. Analyze, characterize and illustrate application of the model.

We start from the simplest possible random variable and build up from there, presenting some results without proof or else leaving such proofs as exercises to the reader where appropriate.

---

## 8.2 The Discrete Uniform Random Variable

### 8.2.1 Basic Characteristics and Model

The phenomenon underlying the discrete uniform random variable is as follows:

1. The experiment has  $k$  mutually exclusive outcomes;
2. Each outcome is equiprobable; and
3. The random variable  $X$  assigns the  $k$  distinct values  $x_i; i = 1, 2, \dots, k$ , to each respective outcome;

The model in this case is quite straightforward:

$$f(x_i) = \begin{cases} \frac{1}{k}; & i = 1, 2, \dots, k; \\ 0; & \text{otherwise} \end{cases} \quad (8.1)$$

with the random variable earning its name because  $f(x)$  is “uniform” across the valid range of admissible values. Thus, Eq (8.1) is the pdf for the discrete Uniform random variable,  $U_D(k)$ . The only characteristic parameter is  $k$ , the total number of elements in the sample space. Sometimes the  $k$  elements are indexed to include 0, i.e.  $i = 0, 1, 2, \dots, k - 1$ , (allowing easier connection to the case where  $k = 2$  and the only two outcomes are the binary numbers 0, 1). Under these circumstances, the mean and variance are:

$$\mu = E(X) = \frac{k}{2} \quad (8.2)$$

and

$$\sigma^2 = \frac{(k+1)(k-1)}{12} \quad (8.3)$$

### 8.2.2 Applications

We have already encountered in Chapters 3 and 4 several examples of the discrete uniform random variable: the tossing of an unbiased coin, (with  $k = 2$ ); the tossing of a fair die (with  $k = 6$  and  $x_i = i; i = 1, 2, \dots, 6$ ); or, in general, the selection of an item at random from a well-defined population of size  $k$  (students from a peer group; balls from a bag; marbles from an urn, etc). This model is therefore most useful in practice for phenomena in which there is no justifiable reason to expect one outcome to be favored over another (See Chapter 10).

In the event that: (i) we restrict the number of outcomes to just two, i.e.  $k = 2$ , (as in the single coin toss; or in the determination of the sex of a newborn selected at random from a hospital; or in an in-vitro fertilization treatment cycle, the success or failure of a single transferred embryo to result in a live birth); and (ii) we relax the equiprobable outcome condition, thereby allowing the probability of the occurrence of each of the two possible outcomes to differ (while necessarily respecting the constraint requiring the two probabilities to sum to 1); the resulting random variable is known as a Bernoulli random variable.

### 8.3 The Bernoulli Random Variable

#### 8.3.1 Basic Characteristics

The phenomenon underlying the Bernoulli random variable is as follows:

1. The experiment has only 2 possible mutually exclusive outcomes, “S” (for success) or “F” (for failure). (Other possible designations are “defective”/“non-defective”; or “Pass”/“Fail”, “Head”/“Tail” etc.);
2. The probability  $p$ , ( $0 < p < 1$ ), is assigned to the outcome “S”;
3. The random variable  $X$  assigns the number 1 to the outcome “S” and 0 to the other outcome, “F”.

A random variable defined as above is a Bernoulli random variable; in fact, an experiment characterized as in item 1 above is known as a “Bernoulli trial.”

#### 8.3.2 Model Development

This is a very straightforward case in which every aspect of the problem is simple and has been specified explicitly. From characteristic 1 above, the sample space is

$$\Omega = \{F, S\} \quad (8.4)$$

consisting of only two elements; and since  $P(S) = p$ , and  $\Omega$  contains exactly two elements,  $P(F)$  must be  $(1-p)$ . Finally, from characteristic 3,  $V_X = \{0, 1\}$  so that:

$$P_X(X = 0) = P(F) = (1 - p) \quad (8.5)$$

$$P_X(X = 1) = P(S) = p \quad (8.6)$$

The desired probability model is therefore given by:

$$f(x) = \begin{cases} (1 - p); & x = 0 \\ p; & x = 1 \end{cases} \quad (8.7)$$

or, in tabular form,

| $x$ | $f(x)$    |
|-----|-----------|
| 0   | $(1 - p)$ |
| 1   | $p$       |

This model can be made more compact as follows: introduce two “indicator” variables, the “success indicator,”  $I_S$ , defined as:

$$I_S = \begin{cases} 1; & \text{for } x = 1 \\ 0; & \text{for } x = 0 \end{cases} \quad (8.8)$$

and its complement, the “failure indicator,”  $I_F$

$$I_F = \begin{cases} 1; & \text{for } x = 0 \\ 0; & \text{for } x = 1 \end{cases} \quad (8.9)$$

The pdf for the Bernoulli random variable is then given by the more compact:

$$f(x) = p^{I_S}(1 - p)^{I_F} \quad (8.10)$$

The Bernoulli random variable,  $Bn(p)$ , is therefore a binary variable; it takes on only two values: 0, with probability  $(1 - p)$ , or 1, with probability  $p$ .

### 8.3.3 Important Mathematical Characteristics

The following are important characteristics of the Bernoulli random variable,  $Bn(p)$ , and its pdf:

1. **Characteristic parameter:**  $p$ ; the probability of “success”.
2. **Mean:**  $\mu = E(X) = p$ .
3. **Variance:**  $\sigma^2 = p(1 - p)$ ; or  $\sigma = \sqrt{p(1 - p)}$
4. **Moment generating function:**  $M(t) = [pe^t + (1 - p)]$
5. **Characteristic function:**  $\varphi(t) = [pe^{jt} + (1 - p)]$

These characteristics are easily established (see Exercise 8.1). For example, by definition,

$$E(X) = \sum_i x_i f(x_i) = 0(1 - p) + 1 \times p = p \quad (8.11)$$

## 8.4 The Hypergeometric Random Variable

### 8.4.1 Basic Characteristics

The hypergeometric random variable arises naturally from problems in acceptance sampling, and similar problems involving drawing samples randomly from a finite-sized population; the basic phenomenon underlying it is as follows:

1. The population (or lot) is dichotomous, in the sense that its elements can be divided into *two* mutually exclusive groups;
2. The total number of units in the lot (equivalently, elements in the population) is  $N$ ;

- $N_d$  of these share a common attribute of interest (e.g. “defective”);
  - the remaining  $(N - N_d)$  do not have this attribute;
  - the population proportion of “defective” items,  $p$ , is therefore  $N_d/N$ ;
3. The experiment: Draw a total of  $n$  items; test them all for the presence of said attribute;
  4. The random variable  $X$ : the number of “defective” items in the sample,  $n$ ;
  5. Assumption: The sample is drawn such that each unit in the lot has an equal chance of being drawn.

#### 8.4.2 Model Development

*The Sample Space:* After each experiment, the outcome  $\omega_i$  is the  $n$ -tuple

$$\omega_i = [a_1, a_2, \dots, a_n]_i \quad (8.12)$$

where each  $a_j; j = 1, 2, \dots, n$  is the attribute of the  $j^{th}$  item drawn ( $a_j$  is therefore either “D” for defective, or “F” for defect-free).

Now the total number of ways of choosing  $n$  items from a lot of size  $N$  is:

$$N_\Omega = \binom{N}{n} = \frac{N!}{n!(N-n)!} \quad (8.13)$$

The sample space, the collection of all such possible outcomes,  $\omega_i$ , therefore contains  $i = 1, 2, \dots, N_\Omega$  elements.

*The Model:* Since the total number of “defectives” contained in each  $\omega_i$  is the random variable of interest in this case, first, observe that obtaining  $x$  “defectives” from a sample size of  $n$  arises from choosing  $x$  from  $N_d$  and  $(n-x)$  from  $(N-N_d)$ ; on the assumption of equiprobable drawings for each item, we obtain

$$P(X = x) = \frac{\text{Total number of “favorable ways” for } X = x}{\text{Total number of possible choices}} \quad (8.14)$$

Thus,

$$f(x) = \frac{\binom{N_d}{x} \binom{N-N_d}{n-x}}{\binom{N}{n}} \quad (8.15)$$

is the pdf for a hypergeometric random variable,  $H(n, N_d, N)$ . Alternatively,

$$f(x) = \frac{1}{c} \frac{N_d!}{x!(N_d-x)!} \frac{(N-N_d)!}{(n-x)!(N-N_d-n+x)!} \quad (8.16)$$

where the constant  $c$  is  $\binom{N}{n}$ .

### 8.4.3 Important Mathematical Characteristics

The following are important characteristics of the hypergeometric random variable,  $H(n, N_d, N)$ , and its pdf:

1. **Characteristic parameters:**  $n, N_d, N$ , respectively, the sample size, total number of “defectives”, and the population or lot size.
2. **Mean:**  $\mu = E(X) = nN_d/N = np$
3. **Variance:**  $\sigma^2 = np(1 - p)\frac{(N-n)}{(N-1)}$ ;

### 8.4.4 Applications

This random variable and its pdf model find application mostly in acceptance sampling. The following are a few examples of such applications.

**Example 8.1 APPLICATION OF THE HYPERGEOMETRIC MODEL**

A batch of 20 electronic chips contains 5 defectives. Find the probability that out of 10 selected for inspection (without replacement) 2 will be found defective.

**Solution:**

In this case,  $x = 2, N_d = 5, N = 20, n = 10$  and therefore:

$$f(x) = 0.348 \quad (8.17)$$

**Example 8.2 APPLICATION OF THE HYPERGEOMETRIC MODEL**

An order of 25 high-reliability electron tubes has been shipped to your company. Your acceptance sampling protocol calls for selecting 5 at random to put through a destructive “accelerated life-test;” if fewer than 2 fail the test, the remaining 20 are accepted, otherwise the shipment is rejected. What is the probability of accepting the lot if truly 4 out of the 25 tubes are defective?

**Solution:**

In this case,  $N_d = 4, N = 25, n = 5$ ; and we require  $P(X = 0, 1)$ :

$$P(\text{Lot acceptance}) = f(0) + f(1) = 0.834 \quad (8.18)$$

so that there is a surprisingly high probability that the lot will be accepted even though 16% is defective. Perhaps the acceptance sampling protocol needs to be re-examined.

## 8.5 The Binomial Random Variable

### 8.5.1 Basic Characteristics

The basic phenomenon underlying the binomial random variable is as follows:

1. Each experiment consists of  $n$  independent “Bernoulli trials” under identical conditions; i.e. each trial produces exactly two mutually exclusive outcomes nominally labeled “S” (success) and “F” (failure);
2. The probability of “success” in each trial,  $P(S) = p$ ;
3. The random variable,  $X$ , is the number of “successes” in the  $n$  trials.

It should be clear from these characteristics that the binomial random variable is related to the Bernoulli random variable and also to the hypergeometric random variable. We will identify the explicit relationships shortly.

More importantly, a good number of real-life problems can be idealized in the form of the simple conceptual experiment of tossing a coin  $n$  times and observing  $x$ , the total number of heads: for example, in *in-vitro* fertilization (IVF) procedures, the total number of live births resulting from the transfer of  $n$  embryos in a patient’s womb, given that  $p$  is the probability of a successful pregnancy from a single embryo, is a binomial random variable (See case study in Chapter 11). Similarly, in characterizing the reliability of a system consisting of  $n$  components, given that  $p$  is the probability of a single component functioning, then  $x$ , the total number of components functioning at any specific time is also a binomial random variable. The binomial random variable is therefore important in its own right; but, as we will soon see, along with its probability model, it also serves as the launching pad for the development of other probability models of important phenomena.

### 8.5.2 Model Development

*The Sample Space:* Each outcome of a “binomial” experiment may be represented as the  $n$ -tuple:

$$\omega_i = [s_1, s_2, \dots, s_n]_i \quad (8.19)$$

a string of  $n$  letters that are either “S” or “F”. Because each trial results in exactly one of two mutually exclusive results (“S” or “F”), there are precisely  $2^n$  of such elements  $\omega_i$  in the sample space  $\Omega$  (recall Example 3.8 in Chapter 3, especially, Eqn. (3.28)); i.e.

$$\Omega = \{w_i\}_1^{2^n} \quad (8.20)$$

*The Random Variable  $X$ :* The total number of occurrences of “S” contained in each  $\omega_i$  is the random variable of interest in this case. The most primitive elementary events, the observation of an “S” or an “F” in each trial, are mutually exclusive; and the observation of a total number of  $x$  occurrences of “S” in each experiment corresponds to the compound event  $E_x$  where

$$E_x = \{x \text{ occurrences of “S” and } (n - x) \text{ occurrences of “F”}\} \quad (8.21)$$

*The Model:* Given that the probability of “success” in each trial,  $P(S) = p$ , and by default  $P(F) = (1 - p) = q$ , then by the independence of the  $n$  trials in each experiment, the probability of the occurrence of the compound event  $E_x$  defined above is:

$$P(E_x) = p^x(1 - p)^{n-x} \quad (8.22)$$

However, in the original sample space  $\Omega$ , there are  $\binom{n}{x}$  different such events in which the sequence in  $\omega_i$  contains  $x$  “successes” and  $(n - x)$  “failures,” where

$$\binom{n}{x} = \frac{n!}{x!(n - x)!} \quad (8.23)$$

Thus,  $P(X = x)$  is the sum of all events contributing to the “pre-image” in  $\Omega$  of the event that the random variable  $X$  takes on the value  $x$ ; i.e.

$$P(X = x) = \binom{n}{x} p^x(1 - p)^{n-x} \quad (8.24)$$

Thus, the pdf for the binomial random variable,  $Bi(n, p)$ , is:

$$f(x) = \frac{n!}{x!(n - x)!} p^x(1 - p)^{n-x} \quad (8.25)$$

### 8.5.3 Important Mathematical Characteristics

The following are important characteristics of the binomial random variable,  $Bi(n, p)$  and its pdf:

1. **Characteristic parameters:**  $n, p$ ; respectively, the number of independent trials in each experiment, and the probability of “success” in each trial;
2. **Mean:**  $\mu = E(X) = np$ ;
3. **Variance:**  $\sigma^2 = np(1 - p)$ ;
4. **Moment generating function:**  $M(t) = [pe^t + (1 - p)]^n$ ;
5. **Characteristic function:**  $\varphi(t) = [pe^{jt} + (1 - p)]^n$ ;

### Relation to other random variables

The binomial random variable is intimately related to other random variables we are yet to encounter (see later); it is also related as follows to the two random variables we have discussed thus far:

**1. Hypergeometric:** It is easy to show that the hypergeometric random variable,  $H(n, N_d, N)$ , and the binomial random variable  $Bi(n, p)$  are related as follows:

$$\lim_{N \rightarrow \infty} H(n, N_d, N) = Bi(n, p) \quad (8.26)$$

where  $p = N_d/N$  remains constant.

**2. Bernoulli:** If  $X_1, X_2, \dots, X_n$  are  $n$  independent Bernoulli random variables, then:

$$X = \sum_{i=1}^n X_i \quad (8.27)$$

is a binomial random variable. This is most easily established by computing the characteristic function for  $X$  as defined in (8.27) from the characteristic function of the Bernoulli random variable. (See Exercise 8.5.)

#### 8.5.4 Applications

The following are a few examples of practical applications of the binomial model.

##### Example 8.3 APPLICATION OF THE BINOMIAL MODEL: ANALYSIS

From experience with a battery manufacturing process, it is known that 5% of the products from a particular site are defective. Find the probability of obtaining 2 defective batteries in a batch of 20 drawn from a large lot manufactured at this particular site, and show that we are more likely to find 1 defective battery in the sample of 20 than 2.

##### Solution:

This problem may be idealized as that of computing  $P(X = 2)$  where  $X$  is a binomial random variable  $Bi(20, 0.05)$  i.e. where  $n$  the number of trials is 20, and the probability of “success” is 0.05. In this case then:

$$P(X = 2) = f(2) = \binom{20}{2} (0.05)^2 (0.95)^{18} = 0.189 \quad (8.28)$$

On the other hand,

$$P(X = 1) = f(1) = \binom{20}{1} (0.05)(0.95)^{19} = 0.377 \quad (8.29)$$

so that we are almost twice as likely to find only 1 defective battery in the sample of 20 than 2.

**Example 8.4 APPLICATION OF THE BINOMIAL MODEL:  
DESIGN**

From the sales record of an analytical equipment manufacturing company, it is known that their sales reps typically make *on average* one sale of a top-of-the-line near-infrared device for every 3 attempts. In preparing a training manual for future sales reps, the company would like to specify  $n$ , the *smallest* number of sales attempts each sales rep should make (per week) such that the probability of scoring an actual sale (per week) is greater than 0.8. Find  $n$ .

**Solution:**

This problem may also be idealized as involving a binomial  $Bi(n, p)$  random variable in which  $p = 1/3$  but for which  $n$  is an unknown to be determined to satisfy a “design” criterion. Finding the probability of the event of interest, ( $X \geq 1$ ), is easier if we consider the complement—the event of making no sale at all ( $X = 0$ ), i.e.

$$P(X \geq 1) = 1 - P(X = 0) \quad (8.30)$$

In this case, since

$$f(0) = \left(\frac{2}{3}\right)^n \quad (8.31)$$

then, we want

$$1 - \left(\frac{2}{3}\right)^n > 0.8 \quad (8.32)$$

from where we obtain the smallest  $n$  to satisfy this inequality to be 4.

Thus, the sales brochure should recommend that each sales rep make at least 4 sales attempt to meet the company goals.

### Inference

A fundamental question about binomial random variables, and indeed all random variables, centers around how the parameters indicated in the pdfs may be determined from data. This is a question that will be considered later in greater detail and in a broader context; for now, we consider the following specific question as an illustration: *Given data, what can we say about  $p$ ?*

In the particular case of a coin toss, this is a question about determining the “true” probability of obtaining a “head” (or “tail”) given data from actual coin toss experiments; in the case of “predicting” the sex of babies, it is about determining the probability of having a boy or girl given hospital birth records; and, as discussed extensively in Chapter 11, in in-vitro fertilization, it is determining from appropriate fertility clinic data, the probability that a particular single embryo will lead to a successful pregnancy. The answer to this specific question is one of a handful of important fundamental results of probability theory.

Let  $X$  be the random variable representing the number of successes in  $n$  independent trials, each with an equal probability of success,  $p$ , so that  $X/n$  is the relative frequency of success.

Since it seems intuitive that the relative frequency of success should be a “reasonable estimate” of the true probability of success, we are interested in computing  $P(|\frac{X}{n} - p| \geq \epsilon)$  for some  $\epsilon > 0$ , i.e. in words,

*what is the probability that the relative frequency of success will differ from the true probability by more than an arbitrarily small number,  $\epsilon$ ?*

Alternatively, this may be restated as:

$$P(|X - np| \geq n\epsilon) \quad (8.33)$$

Cast in terms of Chebyshev’s inequality, which we recall as:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}, \quad (8.34)$$

the bound on the probability we seek is given by:

$$P(|X - np| \geq n\epsilon) \leq \frac{\sigma^2}{\epsilon^2 n^2} \quad (8.35)$$

since, by comparison of the left hand sides of these two equations, we obtain  $k\sigma = n\epsilon$  from which  $k$  is easily determined, giving rise to the RHS of the inequality above. And now because we are particularly concerned with the binomial random variable for which  $\mu = np$  and  $\sigma^2 = np(1 - p)$ , we have:

$$P(|X - np| \geq n\epsilon) \leq \frac{p(1 - p)}{n\epsilon^2} \quad (8.36)$$

For every  $\epsilon > 0$ , as  $n \rightarrow \infty$ ,

$$\lim_{n \rightarrow \infty} \left\{ \frac{p(1 - p)}{n\epsilon^2} \right\} = 0 \quad (8.37)$$

giving the important result:

$$\lim_{n \rightarrow \infty} P \left( \left| \frac{X}{n} - p \right| \geq \epsilon \right) = 0 \quad (8.38)$$

or the complementary result:

$$\lim_{n \rightarrow \infty} P \left( \left| \frac{X}{n} - p \right| \leq \epsilon \right) = 1 \quad (8.39)$$

Together, these two equations constitute one form of the “Law of Large Numbers” indicating, in this particular case, that the relative frequency of success

(the number of successes observed per  $n$  trials) approaches the actual probability of success,  $p$ , as  $n \rightarrow \infty$ , *with probability* 1. Thus, for a large number of trials:

$$\frac{x}{n} \approx p. \quad (8.40)$$

## 8.6 Extensions and Special Cases of the Binomial Random Variable

We now consider a series of random variables that are either direct extensions of the binomial random variable (the trinomial and general multinomial random variables), or are special cases (the negative binomial and the geometric random variables).

### 8.6.1 Trinomial Random Variable

#### Basic Characteristics

In direct analogy to the binomial random variable, the following basic phenomenon underlies the trinomial random variable:

1. Each experiment consists of  $n$  independent trials under identical conditions;
2. Each trial produces exactly *three* mutually exclusive outcomes,  $\omega_1, \omega_2, \omega_3$ , (“Good”, “Average”, “Poor”; A, B, C; etc);
3. In each single trial, the probability of obtaining outcome  $\omega_1$  is  $p_1$ ; the probability of obtaining  $\omega_2$  is  $p_2$ ; and the probability of obtaining  $\omega_3$  is  $p_3$ , with  $p_1 + p_2 + p_3 = 1$ ;
4. The random variable of interest is the two-dimensional, ordered pair  $(X_1, X_2)$ , where  $X_1$  is the number of times that outcome 1,  $\omega_1$ , occurs in the  $n$  trials; and  $X_2$  is the number of times that outcome 2,  $\omega_2$ , occurs in the  $n$  trials. (The third random variable,  $X_3$ , the complementary number of times that outcome 3,  $\omega_3$ , occurs in the  $n$  trials, is constrained to be given by  $X_3 = n - X_1 - X_2$ ; it is not independent.)

#### The Model

It is easy to show, following the same arguments employed in deriving the binomial model, that the trinomial model is:

$$f(x_1, x_2) = \frac{n!}{x_1! x_2! (n - x_1 - x_2)!} p_1^{x_1} p_2^{x_2} p_3^{n-x_1-x_2}, \quad (8.41)$$

the joint pdf for the two-dimensional random variable  $(X_1, X_2)$ .

### Some Important Results

The moment generation function (mgf) for the trinomial random variable is:

$$M(t_1, t_2) = \sum_{x_1=0}^n \sum_{x_2=0}^{n-x_1} \frac{n!}{x_1! x_2! (n - x_1 - x_2)!} (p_1 e^{t_1})^{x_1} (p_2 e^{t_2})^{x_2} p_3^{n-x_1-x_2} \quad (8.42)$$

which simplifies to:

$$M(t_1, t_2) = (p_1 e^{t_1} + p_2 e^{t_2} + p_3)^n \quad (8.43)$$

We now note the following important results:

$$M(t_1, 0) = [(1 - p_1) + p_1 e^{t_1}]^n \quad (8.44)$$

$$M(0, t_2) = [(1 - p_2) + p_2 e^{t_2}]^n \quad (8.45)$$

indicating marginal mgfs, which, when compared with the mgf obtained earlier for the binomial random variable, shows that:

1. The marginal distribution of  $X_1$  is that of the  $Bi(n, p_1)$  binomial random variable;
2. The marginal distribution of  $X_2$  is that of the  $Bi(n, p_2)$  binomial random variable.

### 8.6.2 Multinomial Random Variable

It is now a straightforward exercise to extend the discussion above to the multinomial case where there are  $k$  mutually exclusive outcomes in each trial, each with the respective probabilities of occurrence,  $p_i; i = 1, 2, 3, \dots, k$ , such that:

$$\sum_{i=1}^k p_i = 1 \quad (8.46)$$

The resulting model is the pdf:

$$f(x_1, x_2, \dots, x_k) = \frac{n!}{x_1! x_2! \dots x_k!} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k} \quad (8.47)$$

along with

$$\sum_{i=1}^k x_i = n \quad (8.48)$$

### 8.6.3 Negative Binomial Random Variable

#### Basic Characteristics

The basic phenomenon underlying the negative binomial random variable is very similar to that for the original binomial random variable dealt with earlier:

1. Like the binomial random variable, each trial produces exactly two mutually exclusive outcomes “S” (success) and “F” (failure); the probability of “success” in each trial,  $P(S) = p$ ;
2. The experiment consists of *as many trials as are needed to obtain k successes*, with each trial considered independent, and carried out under identical conditions;
3. The random variable,  $X$ , is the number of “failures” before the  $k^{th}$  “success”. (Since the labels “S” and “F” are arbitrary, this could also be considered as the number of “successes” before the  $k^{th}$  “failure” if it is more logical to consider the problem in this fashion, so long as we are consistent with what we refer to as a “success” and its complement that is referred to as the “failure”.)

#### Model Development

From the definition of the random variable,  $X$ ,  $n$ , the total number of trials required to obtain exactly  $k$  successes is  $X + k$ ; and mechanistically, the event  $X = x$  occurs as a combination of two independent events: (i) obtaining  $x$  failures *and*  $k - 1$  successes in the first  $x + k - 1$  trials *and* (ii) obtaining a success in the  $(x + k)^{th}$  trial. Thus:

$$\begin{aligned} P(X = x) &= P(x \text{ failures and } k - 1 \text{ successes in the first } x + k - 1 \text{ trials}) \\ &\quad \times P(\text{successes in the } (x + k)^{th} \text{ trial}) \end{aligned} \quad (8.49)$$

and from the binomial pdf, we obtain:

$$\begin{aligned} P(X = x) &= \left\{ \binom{x+k-1}{k-1} p^{k-1} (1-p)^x \right\} \times p \\ &= \binom{x+k-1}{k-1} p^k (1-p)^x \end{aligned} \quad (8.50)$$

Thus, the model for the negative binomial random variable  $NBi(k, p)$  is:

$$f(x) = \binom{x+k-1}{k-1} p^k (1-p)^x ; x = 0, 1, 2, \dots \quad (8.51)$$

which is also sometimes written in the entirely equivalent form (see Exercise 8.10):

$$f(x) = \binom{x+k-1}{x} p^k (1-p)^x ; x = 0, 1, 2, \dots \quad (8.52)$$

(In some instances, the random variable is defined as the total number of *trials* required to obtain exactly  $k$  “successes”; the discussion above is easily modified for such a definition of  $X$ . See Exercise 8.10).

In the most general sense, the parameter  $k$  of the negative binomial random variable in fact does *not* have to be an integer. In most engineering applications, however,  $k$  is almost always an integer. In honor of the French mathematician and philosopher, Blaise Pascal (1623–1662), in whose work one will find the earliest mention of this distribution, the negative binomial distribution with integer  $k$  is often called the Pascal distribution. When the parameter  $k$  is real-valued, the pdf is known as the Polya distribution, in honor of the Hungarian mathematician, George Pólya (1887–1985), and written as:

$$f(x) = \frac{\Gamma(x+k)}{\Gamma(k)x!} p^k (1-p)^x ; x = 0, 1, 2, \dots \quad (8.53)$$

where  $\Gamma(\alpha)$  is the Gamma function defined as:

$$\Gamma(\alpha) = \int_0^\infty e^{-y} y^{\alpha-1} dy \quad (8.54)$$

This function satisfies the recursive expression:

$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha) \quad (8.55)$$

so that if  $\alpha$  is a positive integer, then

$$\Gamma(\alpha) = (\alpha - 1)! \quad (8.56)$$

and the pdf in Eqs (8.53) will coincide with that in Eq (8.51) or Eq (8.52).

### Important Mathematical Characteristics

The following are important characteristics of the negative binomial random variable,  $NBi(k, p)$ , and its pdf:

1. **Characteristic parameters:**  $k, p$ ; respectively, the target number of successes, and the probability of “success” in each trial;
2. **Mean:**  $\mu = E(X) = \frac{k(1-p)}{p} = \frac{kq}{p}$ ;
3. **Variance:**  $\sigma^2 = \frac{k(1-p)}{p^2} = \frac{kq}{p^2}$ ;
4. **Moment generating function:**  $M(t) = p^k (1 - qe^t)^{-k}$
5. **Characteristic function:**  $\varphi(t) = p^k (1 - qe^{jt})^{-k}$ .

An alternative form of the negative binomial pdf arises from the following re-parameterization: Let the mean value be represented as  $\lambda$ , i.e.,

$$\lambda = \frac{k(1-p)}{p} = k \left( \frac{1}{p} - 1 \right) \quad (8.57)$$

so that

$$p = \frac{k}{k + \lambda} \quad (8.58)$$

in which case, Eq (8.51) becomes

$$f(x) = \frac{(x+k-1)!}{(k-1)!x!} \frac{\lambda^x}{\left(1 + \frac{\lambda}{k}\right)^k} \quad (8.59)$$

We shall have cause to refer to this form of the pdf shortly.

#### 8.6.4 Geometric Random Variable

Consider the special case of the negative binomial random variable with  $k = 1$ ; where the resulting random variable  $X$  is the number of “failures” before the first success. It follows immediately from Eqn. (8.51) that the required pdf in this case is:

$$f(x) = pq^x; \quad x = 0, 1, 2, \dots \quad (8.60)$$

#### The Model

It is more common to consider the geometric random variable as the number of *trials*—not “failures”—required to obtain the first success. It is easy to see that this definition of the geometric random variable merely requires a shift by one in the random variable discussed above, so that the pdf for the geometric random variable is given by:

$$f(x) = pq^{x-1}; \quad x = 1, 2, \dots \quad (8.61)$$

#### Important Mathematical Characteristics

The following are important characteristics of the geometric random variable,  $G(p)$ , and its pdf:

1. **Characteristic parameter:**  $p$ , the probability of “success” in each trial;
2. **Mean:**  $\mu = E(X) = \frac{1}{p}$ ;
3. **Variance:**  $\sigma^2 = \frac{q}{p^2}$ ;
4. **Moment generating function:**  $M(t) = \frac{p}{q}(1 - qe^t)^{-1}$
5. **Characteristic function:**  $\varphi(t) = \frac{p}{q}(1 - qe^{jt})^{-1}$ .

## Applications

One of the most important applications of the geometric random variable is in free radical polymerization where, upon initiation, monomer units add to a growing chain, with each subsequent addition propagating the chain until a termination event stops the growth. After initiation, each “trial” involves either a “propagation” event (the successful addition of a monomer unit to the growing chain), or a “termination” event, where the polymer chain is “capped” to yield a “dead” polymer chain that can no longer add another monomer unit. Because the outcome of each “trial” (propagation or termination) is random, the resulting polymer chains are of variable length; in fact, the physical properties and performance characteristics of the polymer are related directly to the chain length distribution. It is therefore of primary interest to characterize polymer chain length distributions appropriately.

Observe that as described above, the phenomenon underlying free-radical polymerization is such that each polymer chain length is precisely the total number of monomer units added until the occurrence of the termination event. Thus, if termination is considered a “success,” then the chain length is a geometric random variable. In polymer science textbooks (e.g. Williams, 1971<sup>1</sup>), chemical kinetics arguments are often used to establish what is referred to as the “most probable chain length distribution;” the result is precisely the geometric pdf presented here. In Chapter 10, we use maximum entropy considerations to arrive at the same results.

### Example 8.5 APPLICATION OF THE GEOMETRIC DISTRIBUTION MODEL

From their prior history, it is known that the probability of a building construction company recording an accident (minor or major) on any day during construction is 0.2. (a) Find the probability of going 7 days (since the last occurrence) before recording the 1<sup>st</sup> accident. (b) What is the *expected* number of days before recording the 1<sup>st</sup> accident?

**Solution:**

This problem clearly involves the geometric random variable with  $p = 0.2$ . Thus

(a) the required  $P(X = 7)$  is obtained as:

$$f(7) = 0.2(0.8)^6 = 0.05 \quad (8.62)$$

so that because of the relatively high probability of the occurrence of an accident, it is highly unlikely that this company can go 7 days before recording the first accident.

(b) The expected value of days in between accidents is:

$$E(X) = \frac{1}{0.2} = 5 \text{ days} \quad (8.63)$$

so that one would expect on average 5 days before recording an accident.

---

<sup>1</sup>D.J. Williams, *Polymer Science and Engineering*, Prentice Hall, NJ, 1971, pp58-59

## 8.7 The Poisson Random Variable

The Poisson random variable is encountered in so many practical applications, ranging from industrial manufacturing to physics and biology, and even in such military problems as the historic study of deaths by horse-kicks in the Prussian army, and the German bombardment of London during World War II.

We present here two approaches to the development of the probability model for this important random variable: (i) as a limiting form of the binomial (*and* negative binomial) random variable; (ii) from “first principles.”

### 8.7.1 The Limiting Form of a Binomial Random Variable

Consider a binomial random variable under the following conditions:

1. The number of trials is very large ( $n \rightarrow \infty$ );
2. But as the number of trials becomes very large, the probability of success becomes very small to the same proportion such that  $np = \lambda$  remains constant.

i.e. we wish to consider the binomial random variable in the limit as  $n \rightarrow \infty$  but with  $np$  remaining constant at the value  $\lambda$ . The underlying phenomenon is therefore that of the occurrence of rare events (with very small probabilities of occurrence) in a large number of trials.

#### Model Development

From Eq. (11.1), the pdf we seek is given by:

$$f(x) = \lim_{\substack{n \rightarrow \infty \\ np = \lambda}} \left\{ \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \right\} \quad (8.64)$$

which may be written as:

$$\begin{aligned} f(x) &= \lim_{n \rightarrow \infty} \left\{ \frac{n(n-1)(n-2)\dots(n-x+1)(n-x)!}{x!(n-x)!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \right\} \\ &= \lim_{n \rightarrow \infty} \left\{ \frac{1(1-1/n)(1-2/n)\dots(1-(x-1)/n)}{x!} \lambda^x \frac{\left(1 - \frac{\lambda}{n}\right)^n}{\left(1 - \frac{\lambda}{n}\right)^x} \right\} \quad (8.65) \end{aligned}$$

Now, because:

$$\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda} \quad (8.66)$$

$$\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^x = 1 \quad (8.67)$$

the latter being the case because  $x$  is fixed,  $f(x)$  therefore reduces to:

$$f(x) = \frac{e^{-\lambda} \lambda^x}{x!}; \quad x = 0, 1, 2, \dots \quad (8.68)$$

This is the pdf of the Poisson random variable,  $\mathcal{P}(\lambda)$ , with the parameter  $\lambda$ .

It is also straightforward to show (see Exercise 8.14), that the Poisson pdf arises in the limit as  $k \rightarrow \infty$  for the negative binomial random variable, but with the mean  $kq/p = \lambda$  remaining constant; i.e., from Eq (8.59),

$$f(x) = \lim_{k \rightarrow \infty} \left\{ \frac{(x+k-1)!}{(k-1)!x!} \frac{\lambda^x}{\left(1 + \frac{\lambda}{k}\right)^k} \right\} = \frac{e^{-\lambda} \lambda^x}{x!}; \quad x = 0, 1, 2, \dots \quad (8.69)$$

### 8.7.2 First Principles Derivation

#### Basic Characteristics

Considered from “first principles,” the basic phenomenon underlying the Poisson random variable is as follows:

1. The experiment consists of observing the number of occurrences of a particular event (a “success”) in a given fixed interval (of time, length, space) or area, or volume, etc, of size  $z$ ;
2. The probability,  $p$ , of observing exactly one “success” in a sub-interval of size  $\Delta z$  units (where  $\Delta z$  is small), is proportional to  $\Delta z$ ; i.e.

$$p = \eta \Delta z \quad (8.70)$$

Here,  $\eta$ , the rate of occurrence of the “success” per unit interval, is constant. The probability of *not* observing the “success” is  $(1 - p)$ .

3. The probability of observing more than 1 “success” within the sub-interval is negligible, by choice of  $\Delta z$ . (The mathematical implication of this statement is that the indicated probability is  $\mathcal{O}(\Delta z)$ , a quantity that goes to zero faster than  $\Delta z$ , such that  $\lim_{\Delta z \rightarrow 0} \{\mathcal{O}(\Delta z)/\Delta z\} = 0$ );
4. *Homogeneity*: The location of any particular sub-interval of size  $\Delta z$  is immaterial and does not affect the probability of “success” or “failure”;
5. *Independence*: The occurrence of a “success” in one sub-interval is independent of the occurrence in any other sub-interval;
6. The random variable,  $X$ , is the total number of “successes” observed in the entire interval of size  $z$ .

Observe that these characteristics fit that of a binomial random variable with a large number of “trials,”  $n = (z/\Delta z)$ , each trial having two mutually exclusive outcomes, “success” or “failure”, with a small probability of “success” ( $\eta \Delta z$ ).

### Model Development

We start by defining  $P_x(z)$ ,

$$P_x(z) = P(X = x \text{ in an interval of size } z) \quad (8.71)$$

Then the probability of observing  $x$  “successes” in an interval of size  $z + \Delta z$  is given by

$$P_x(z + \Delta z) = P(E_1 \cup E_2 \cup E_3) \quad (8.72)$$

where  $E_1, E_2$  and  $E_3$  are mutually exclusive events defined as follows:

- $E_1$ : exactly  $x$  “successes” are observed in the interval of size  $z$  and none are observed in the adjacent sub-interval of size  $\Delta z$ ;
- $E_2$ : exactly  $x - 1$  “successes” are observed in the interval of size  $z$  and exactly one more is observed in the adjacent sub-interval of size  $\Delta z$ ;
- $E_3$ : exactly  $x - i$  “successes” are observed in the interval of size  $z$  and exactly  $i$  more are observed in the adjacent sub-interval of size  $\Delta z$ , with  $i = 2, 3, \dots, x$ ;

From the phenomenological description above, we have the following results:

$$P(E_1) = P_x(z)(1 - \eta\Delta z) \quad (8.73)$$

$$P(E_2) = P_{x-1}(z)\eta\Delta z \quad (8.74)$$

$$P(E_3) = \mathcal{O}(\Delta z) \quad (8.75)$$

Hence,

$$P_x(z + \Delta z) = P_x(z)(1 - \eta\Delta z) + P_{x-1}(z)\eta\Delta z + \mathcal{O}(\Delta z); x = 1, 2, \dots \quad (8.76)$$

In particular, for  $x = 0$ , we have:

$$P_0(z + \Delta z) = P_0(z)(1 - \eta\Delta z) \quad (8.77)$$

since in this case, both  $E_2$  and  $E_3$  are impossible events. Dividing by  $\Delta z$  in both Eqs (8.76) and (8.77) and rearranging gives

$$\frac{P_x(z + \Delta z) - P_x(z)}{\Delta z} = -\eta[P_x(z) - P_{x-1}(z)] + \frac{\mathcal{O}(\Delta z)}{\Delta z} \quad (8.78)$$

$$\frac{P_0(z + \Delta z) - P_0(z)}{\Delta z} = -\eta P_0(z) \quad (8.79)$$

from where taking limits as  $\Delta z \rightarrow 0$  produces the following series of differential equations:

$$\frac{dP_x(z)}{dz} = -\eta[P_x(z) - P_{x-1}(z)] \quad (8.80)$$

$$\frac{dP_0(z)}{dz} = -\eta P_0(z) \quad (8.81)$$

To solve these equations requires the following initial conditions:  $P_0(0)$ , the probability of finding no “success” in the interval of size 0 — a certain event — is 1;  $P_x(0)$ , the probability of finding  $x$  “successes” in the interval of size 0 — an impossible event — is 0. With these initial conditions, we obtain, first for  $P_0(z)$ , that

$$P_0(z) = e^{-\eta z} \quad (8.82)$$

which we may now introduce into Eq. (8.80) and solve recursively for  $x = 1, 2, \dots$  to obtain, in general (after some tidying up),

$$P_x(z) = \frac{(\eta z)^x e^{-(\eta z)}}{x!} \quad (8.83)$$

Thus, from first principles, the model for the Poisson random variable is given by:

$$f(x) = \frac{(\eta z)^x e^{-\eta z}}{x!} = \frac{\lambda^x e^{-\lambda}}{x!} \quad (8.84)$$

### 8.7.3 Important Mathematical Characteristics

The following are important characteristics of the Poisson random variable,  $\mathcal{P}(\lambda)$ , and its pdf:

1. **Characteristic parameters:**  $\lambda$ , or  $\eta z$ , the mean number of “successes” in the entire interval of size  $z$ ; and  $\eta$ , the mean number of “successes” per unit interval size, is sometimes called the “intensity”.
2. **Mean:**  $\mu = E(X) = \lambda$
3. **Variance:**  $\sigma^2 = \lambda$ ;
4. **Moment generating function:**  $M(t) = e^{[\lambda(e^t - 1)]}$ ;
5. **Characteristic function:**  $\varphi(t) = e^{[\lambda(e^{jt} - 1)]}$
6. **Reproductive Properties:** The Poisson random variable (as with a few others to be discussed later) possesses the following useful property: If  $X_i, i = 1, 2, \dots, n$ , are  $n$  independent Poisson random variables each with parameter  $\lambda_i$ , i.e.  $X_i \sim \mathcal{P}(\lambda_i)$ , then the random variable  $Y$  defined as:

$$Y = \sum_{i=1}^n X_i \quad (8.85)$$

is also a Poisson random variable, with parameter  $\lambda^* = \sum_{i=1}^n \lambda_i$ . Because a sum of Poisson random variables begets another Poisson random variable, this characteristic is known as a “reproductive” property. This result is easily established using the method of characteristic functions discussed in Chapter 6. (See Exercise 8.17.)

### 8.7.4 Applications

#### Standard Poisson Phenomena

This Poisson random variable and its pdf find application in a wide variety of practical problems. Recall, for example, the problem of “Quality Assurance in a Glass Sheet Manufacturing Process” considered in Chapter 1, involving the number of “inclusions” per square meter of a manufactured glass sheet. The pdf we simply stated in that chapter for the random variable in question—without justification—is now recognizable as the Poisson pdf. The same is true of the application to the number of cell divisions in a fixed time interval  $t$  discussed in Chapter 6. Following the preceding discussion, it should now be clear to the reader why this was in fact the appropriate pdf to use in each of these applications: the random variable of concern in each example (the number of “inclusions” in the Chapter 1 example; the number of times each cell in the cell culture divides in the Chapter 6 application) each possesses all the characteristics noted above for the Poisson random variable.

Not surprisingly, the Poisson model also finds application in the analysis of annual/monthly/weekly occurrences of safety incidents in manufacturing sites; the number of yarn breaks per shift in fiber spinning machines, and other such phenomena involving counts of occurrences of rare events in a finite interval. The pdf is also used as an approximation to binomial random variables with large  $n$  and small  $p$ , (with  $np \leq 7$ ), where the binomial pdf would have been quite tedious to use. The following are a few illustrative examples:

#### Example 8.6 APPLICATION OF THE POISSON DISTRIBUTION MODEL

Silicon wafers of a particular size made by a chip manufacturer are known to have an average of two contaminant particles each. Determine the probability of finding more than 2 contaminant particles on any such wafer chosen at random.

##### Solution:

This problem involves a Poisson random variable with  $\lambda = 2$ . Thus, since

$$P(X > 2) = (1 - P(X \leq 2)) \quad (8.86)$$

the required probability is obtained as:

$$P(X > 2) = 1 - (f(0) + f(1) + f(2)) \quad (8.87)$$

when  $f(x)$  is given by:

$$f(x) = \frac{e^{-2} 2^x}{x!} \quad (8.88)$$

so that:

$$P(X > 2) = 1 - (0.135 + 0.271 + 0.271) = 0.325 \quad (8.89)$$

#### Example 8.7 APPLICATION OF THE POISSON DISTRIBUTION MODEL

Given that the probability of finding 1 blemish in a foot-long length of

**TABLE 8.1:** Theoretical versus empirical frequencies for *inclusions* data

| $x$ | Theoretical<br>$f(x)$ | Empirical<br>Frequency |
|-----|-----------------------|------------------------|
| 0   | 0.3679                | 0.367                  |
| 1   | 0.3679                | 0.383                  |
| 2   | 0.1839                | 0.183                  |
| 3   | 0.0613                | 0.017                  |
| 4   | 0.0153                | 0.033                  |
| 5   | 0.0031                | 0.017                  |
| 6   | 0.0005                | 0.000                  |

a fiber optics wire is  $1/1000$ , and that the probability of finding more than one blemish in this foot-long unit is  $\approx 0$ , determine the probability of finding 5 blemishes in a 3,000 ft roll of wire.

**Solution:**

This problem involves a Poisson random variable with  $\Delta z = 1$  foot; and the intensity  $\eta = 1/1000$  per foot. For  $z = 3,000$  ft,

$$\lambda = \eta z = 3.0 \quad (8.90)$$

and the required probability is obtained as:

$$f(5) = \frac{e^{-3} 3^5}{5!} = 0.101 \quad (8.91)$$

The next example is a follow up to the illustrative example in Chapter 1.

**Example 8.8 APPLICATION OF THE POISSON MODEL TO QUALITY ASSURANCE IN GLASS MANUFACTURING**

If  $X$ , the number of *inclusions* found on glass sheets produced in the manufacturing process discussed in Chapter 1 can be considered as a Poisson random variable with theoretical value  $\lambda = 1$ , (1) compute the theoretical probabilities of observing  $x = 0, 1, 2, \dots, 6$  inclusions, and compare these with the empirical frequencies generated from the data in Table 1.2 and shown in Table 1.5. (2) If, as stated in Chapter 1, only sheets with 3 or fewer *inclusions* are acceptable and can be sold unconditionally, theoretically what percentage of the product made in this process can be expected to fall into this desired category? (3) What is the theoretical probability of this process producing sheets with 5 or more *inclusions*?

**Solution:**

(1) From Eq. (8.68) with  $\lambda = 1$ , we obtain the values of  $f(x)$  shown in Table 8.1 along with the empirical frequencies computed from the data given in Chapter 1. (We have deliberately included an extra significant figure in the computed  $f(x)$  to facilitate comparison.) Observe how close the theoretical probabilities are to the empirical frequencies, especially for  $x = 0, 1, 2$  and 6. Rigorously and quantitatively determining whether

the discrepancies observed for values of  $x = 3, 4$  and  $5$  are “significant” or not is a matter taken up in Part IV.

(2) The probability of obtaining 3 or fewer *inclusions* is computed as follows:

$$P(X \leq 3) = F(3) = \sum_{x=0}^3 f(x) = 0.981 \quad (8.92)$$

implying that 98.1% of glass sheets manufactured in this process can theoretically be expected to be acceptable, according to the stated criterion of 3 or fewer *inclusions*.

(3) The required probability,  $P(X > 5)$ , is obtained as follows:

$$P(X > 5) = 1 - P(X \leq 5) = 1 - 0.9994 = 0.0006, \quad (8.93)$$

indicating a very small probability that this process will produce sheets with 5 or more *inclusions*.

### Overdispersed Poisson-like Phenomena and the Negative Binomial distribution

The equality of mean and variance is a unique distinguishing characteristic of the Poisson random variable. As such, for *true* Poisson phenomena, it will always be the case that  $\mu = \sigma^2 = \lambda$  in theory, and, within limits of data variability, in practice. There are cases of practical importance, however, where the Poisson-like phenomenon (e.g., the number of occurrences of rare events—with small probabilities of occurrence) possesses a rate (or intensity) parameter that is *not* uniformly constant. Under these circumstances, the variance of the random variable will exceed the mean, giving rise to what is generally known as “overdispersion,” for which the Poisson model will no longer be strictly valid. Examples include such phenomena as counts of certain species of insects found in sectors of a farmland; the number of accidents reported per month to an insurance company; or the number of incidents of suicide per year in various counties in a state. These phenomena clearly involve the occurrences of rare events, but in each case, the characteristic Poisson parameter is not uniform across the entire domain of interest. With the insects, for example, the spatial aggregation per unit area is not likely to be uniform across the entire farm area because certain areas may be more attractive to the insects than others; the susceptibility to accidents is not constant across all age groups; and with human populations, not everyone in the region of interest is subject to the same risk for suicide.

For such problems, the negative binomial model is more appropriate. First, observe from Eq (8.69) that the negative binomial pdf, with two parameters,  $k$  and  $p$ , provides more flexibility (with finite  $k$ ) than the limiting Poisson case; furthermore, the variance,  $k(1 - p)/p^2$ , is *always* different from—in fact *always* larger than—the mean,  $k(1 - p)/p$ , a pre-requisite for overdispersion. More fundamentally, however, it can be shown from first principles that the negative binomial model is in fact *the* appropriate model for such phenomena.

We postpone until Section 9.1 of Chapter 9 the establishment of this result because it requires the use of a probability model that will not be discussed until then. In the meantime, Application Problem 8.28 presents an abbreviated version of the original historical application.

---

## 8.8 Summary and Conclusions

If proper analysis of randomly varying phenomena must begin with an appropriate model of the phenomenon in question, then the importance of this chapter's techniques and results to such analysis cannot be overstated. This chapter is where the generic pdf characterized extensively in Chapters 4 and 5 first begins to take on specific and distinct "personalities" in the form of unique and identifiable probability models for various discrete randomly varying phenomena. In developing these probability models, we began, in each case, with a description of the underlying phenomena; and to turn these descriptions into mathematical models, we invoked, to varying degrees, the ideas of the sample space and probability discussed in Chapter 3, the random variable and pdf of Chapters 4 and 5, and, in a limited number of cases, the techniques of random variable transformations of Chapter 6. The end result has been a wide array of probability models for various discrete random variables, the distinguishing characteristics associated with each model, and the application to which each model is most naturally suited.

The modeling exercises of this chapter have also provided insight into how the various models, and the random phenomena they represent, are related. For example, we now know that the geometric distribution—applicable to problems of chain length distribution in free-radical polymerization (among other applications)—is a special case of the negative binomial distribution, itself a variation of the binomial distribution, which arises from a sum of  $n$  Bernoulli random variables. Similarly, we also know that random variables representing the number of occurrences of rare events in a finite interval of length, area, volume, or time, are mostly Poisson distributed, except when they are "overdispersed," in which case the negative binomial distribution is more appropriate.

Finally, we note that what began in this chapter for discrete random variables continues in the next chapter, employing the same approach, for continuous random variables. In fact, Chapter 9 picks up precisely where this chapter leaves off—with the Poisson random variable. It is advisable, therefore, before engaging with the material in the next chapter, that what has been learned from this chapter be consolidated by tackling as many of the exercises and application problems found at the end of this chapter as possible.

Here, and in Table 8.2, is a summary of the main characteristics, models, and other important features of the discrete random variables of this chapter.

- *Discrete uniform random variable*,  $U_D(k)$ : a variable with  $k$  equiprobable outcomes.
- *Bernoulli random variable*,  $Bn(p)$ : the outcome of a Bernoulli trial—a trial with *only* two mutually exclusive outcomes, 0 or 1; “Success” or “Failure”, etc, with respective probabilities  $p$ , and  $1 - p$ .
- *Hypergeometric random variable*,  $H(n, N_d, N)$ : the number of “defective” items found in a sample of size  $n$  drawn from a population of size  $N$ , of which the total number of “defectives” is  $N_d$ .
- *Binomial random variable*,  $Bi(n, p)$ : the total number of “successes” in  $n$  independent Bernoulli trials with probability of “success,”  $p$ .
- *Multinomial random variable*,  $MN(n, p_1, p_2, \dots, p_k)$ : the total number of times each mutually exclusive outcome  $i$ ;  $i = 1, 2, \dots, k$ , occurs in  $n$  independent trials, with probability of a single occurrence of outcome  $i$ ,  $p_i$ .
- *Negative binomial random variable*,  $NBi(k, p)$ : the total number of “failures” before the  $k^{th}$  “success,” with probability of “success,”  $p$ .
- *Geometric random variable*,  $G(p)$ : the total number of “failures” before the  $1^{st}$  “success,” with probability of “success,”  $p$ .
- *Poisson random variable*,  $\mathcal{P}(\lambda)$ : the total number of (rare) events occurring in a finite interval of length, area, volume, or time, with mean rate of occurrence,  $\lambda$ .

## REVIEW QUESTIONS

1. What are the two primary benefits of the “first principles” approach to probability modeling as advocated in this chapter?
2. What are the four components of the model development and analysis strategy outlined in this chapter?
3. What are the basic characteristics of the discrete uniform random variable?
4. What is the probability model for the discrete uniform random variable?
5. What are some examples of a discrete uniform random variable?
6. What are the basic characteristics of the Bernoulli random variable?
7. What is the connection between the discrete uniform and the Bernoulli random variables?

## Ideal Models of Discrete Random Variables

TABLE 8.2: Summary of probability models for discrete random variables

| Random Variable                | Probability Model   | Characteristic Parameters | Mean ( $\mu$ )                             | Variance ( $\sigma^2$ )   | Relation to Other Variables   |
|--------------------------------|---|---------------------------|--|---|---|
| Uniform $U_D(k)$               | $f(x_i) = \frac{1}{k}; i = 1, 2, \dots, k$<br>or $i = 0, 1, 2, \dots, k - 1$  | $k$<br>$\frac{k+1}{2}$    | $\frac{k+1}{2}$<br>$\frac{(k-1)(k+1)}{12}$ | $Var(X)$  | $U_D(2) = Bi(0.5)$  |
| Bernoulli $Bn(p)$              | $f(x=0) = (1-p)$<br>$f(x=1) = p$  | $p$                       | $pq$<br>$(q=1-p)$                          | $X_i \sim Bi(n, p)$<br>$\sum_{i=1}^n X_i \sim Bi(n, p)$                               |   |
| Hypergeometric $H(n, N_d, N)$  | $f(x) = \frac{\binom{N_d}{x} \binom{N-N_d}{n-x}}{\binom{N}{n}}$   | $n, N_d, N$               | $\frac{nN_d}{N} = np$<br>$(q=1-p)$         | $npq \left( \frac{N-n}{N-1} \right)$<br>$= Bi(n, p)$                                  | $\lim_{N \rightarrow \infty} H(n, N_d, N) = Bi(n, p)$   |
| Binomial $Bi(n, p)$            | $f(x) = \binom{n}{x} p^x (1-p)^{n-x}$<br>$x = 1, 2, \dots, n$   | $n, p$                    | $np$<br>$np(1-p)$                          | $np_i(1-p_i)$<br>$np = \lambda$   | $\lim_{n \rightarrow \infty} Bi(n, p) = \mathcal{P}(\lambda)$   |
| Multinomial $NN(n, p_i)$       | $f(x_1, x_2, \dots, x_k) = \frac{n!}{x_1! x_2! \dots x_k!} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k}$<br>$\sum x_i = n; \sum p_i = 1$ | $n, p_i$                  | $np_i$<br>$np_i(1-p_i)$                    | $f_i(x_i) \sim Bi(n, p_i)$  | Marginal  |
| Negative Binomial $NBi(k, p)$  | $f(x) = \binom{x+k-1}{k-1} p^k (1-p)^x$<br>$x = 1, 2, \dots$  | $k, p$                    | $k(1-p)/p$<br>$k(1-p)/p^2$                 | $NBi(1, p) = G(p);$<br>$\lim_{k \rightarrow \infty} NBi(k, p) = \mathcal{P}(\lambda)$ |   |
| Geometric $G(p)$               | $f(x) = p(1-p)^{x-1}$<br>$x = 1, 2, \dots$  | $p$                       | $1/p$                                      | $(1-p)/p^2$   | $NBi(1, p)$   |
| Poisson $\mathcal{P}(\lambda)$ | $f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$<br>$x = 1, 2, \dots$   | $\lambda$                 | $\lambda$                                  | $\lambda$   | $\lim_{n \rightarrow \infty} Bi(n, p) = \mathcal{P}(\lambda)$<br>$\lim_{k \rightarrow \infty} NBi(k, p) = \mathcal{P}(\lambda)$ |

- 8.** What is a Bernoulli trial?
- 9.** What are the two versions of the probability model for the Bernoulli random variable?
- 10.** What are the basic characteristics of the hypergeometric random variable?
- 11.** What is the probability model for the hypergeometric random variable?
- 12.** What do the parameters,  $n, N_d$  and  $N$  represent for the hypergeometric,  $H(n, N_d, N)$  random variable?
- 13.** What are the basic characteristics of the binomial random variable?
- 14.** What is the probability model for the binomial random variable?
- 15.** What is the relationship between the hypergeometric and binomial random variables?
- 16.** What is the relationship between the Bernoulli and binomial random variables?
- 17.** Chebychev's inequality was used to establish what binomial random variable result?
- 18.** What are the basic characteristics of the trinomial random variable?
- 19.** What is the probability model for the trinomial random variable?
- 20.** What is the relationship between the trinomial and binomial random variables?
- 21.** What is the probability model for the multinomial random variable?
- 22.** What are the basic characteristics of the negative binomial random variable?
- 23.** What is the probability model for the negative binomial random variable?
- 24.** What is the connection between the negative binomial, Pascal and Polya distributions?
- 25.** What is the relationship between the negative binomial and the geometric random variables?
- 26.** What is the probability model for the geometric random variable?
- 27.** The Poisson random variable can be obtained as a limiting case of which random variables, and in what specific ways?
- 28.** What are the basic characteristics of the Poisson random variable?

- 29.** What is the probability model for the Poisson random variable?
- 30.** The Poisson model is most appropriate for what sort of phenomena?
- 31.** What about the mean and the variance of the Poisson random variable is a distinguishing characteristic of this random variable?
- 32.** What is an “overdispersed” Poisson-like phenomena? Give a few examples.
- 33.** What probability model is more appropriate for “overdispersed” Poisson-like phenomena and why?

## EXERCISES

**8.1** Establish the results given in the text for the variance, MGF and CF for the Bernoulli random variable.

**8.2** Given a hypergeometric  $H(n, N_d, N)$  random variable  $X$  for which  $n = 5, N_d = 2$  and  $N = 10$ :

- (i) Determine and plot the entire pdf,  $f(x)$  for  $x = 0, 1, 2, \dots, 5$
- (ii) Determine  $P(X > 1)$  and  $P(X < 2)$
- (iii) Determine  $P(1 \leq X \leq 3)$

**8.3** A crate of 100 apples contains 5 that are rotten. A grocer purchasing the crate selects a sample of 10 apples at random and accepts the entire crate *only* if this sample contains no rotten apples. Determine the probability of accepting the crate. If the sample size is increased to 20, find the new probability of accepting the crate.

**8.4** From the expression for the pdf of a binomial  $Bi(np)$  random variable,  $X$ , establish that  $E(X) = np$  and  $Var(X) = npq$  where  $q = (1 - p)$ .

**8.5** Establish that if  $X_1, X_2, \dots, X_n$  are  $n$  independent Bernoulli random variables, then:

$$X = \sum_{i=1}^n X_i$$

is a binomial random variable.

**8.6** Given that  $X$  is a hypergeometric random variable with  $n = 10, N_d = 5$ , and  $N = 20$ , compute  $f(x)$  for  $x = 0, 1, 2, \dots, 10$ , and compare with the corresponding  $f(x)$  for  $x = 0, 1, 2, \dots, 10$ , for a binomial random variable with  $n = 10, p = 0.25$

**8.7** Obtain the recursion formula

$$f(x+1) = \rho(n, x, p)f(x) \quad (8.94)$$

for the binomial pdf, and show that

$$\rho(n, x, p) = \frac{n-x}{x+1}$$

Use this to determine the value  $x^*$  for which the pdf attains a maximum. (Keep in mind that because  $f(x)$  is *not* a continuous function of  $x$ , the standard calculus approach of finding optima by taking derivatives and setting to zero is invalid. Explore the finite difference  $f(x+1) - f(x)$  instead.)

**8.8** Given the joint pdf for the two-dimensional ordered pair  $(X_1, X_2)$  of the trinomial random variable (see Eq (8.41)), obtain the conditional pdfs  $f(x_1|x_2)$  and  $f(x_2|x_1)$ .

**8.9** Consider a chess player participating in a two-game pre-tournament qualification series. From past records in such games, it is known that the player has a probability  $p_w = 0.75$  of winning, a probability  $p_D = 0.2$  of drawing, and a probability  $p_L = 0.05$  of losing. If  $X_1$  is the number of wins and  $X_2$  is the number of draws, obtain the complete joint pdf  $f(x_1, x_2)$  for this player. From this, compute the marginal pdfs,  $f_1(x_1)$  and  $f_2(x_2)$ , and finally obtain the conditional pdfs  $f(x_1|x_2)$  and  $f(x_2|x_1)$ .

**8.10** (i) Establish the equivalence of Eq (8.51) and Eq (8.52), and also the equivalence of Eq (8.53) and Eq (8.52) when  $k$  is a positive integer.

(ii) If the negative binomial random variable is defined as the total number of *trials* (not “failures”) required to obtain exactly  $k$  “successes,” obtain the probability model in this case and compare it to the model given in Eq (8.51) or Eq (8.52).

**8.11** Obtain the recursion formula

$$f(x+1) = \rho(k, x, p)f(x) \quad (8.95)$$

for the negative binomial pdf, showing an explicit expression for  $\rho(k, x, p)$ . Use this expression to determine the value  $x^*$  for which the pdf attains a maximum. (See comments in Exercise 8.7.) From this expression, confirm that the geometric distribution is monotonically decreasing.

**8.12** (i) Establish that  $E(X)$  for the geometric random variable is  $1/p$  and that  $Var(X) = q/p^2$ , where  $q = 1 - p$ .

(ii) Given that for a certain geometric random variable,  $P(X = 2) = 0.0475$  and  $P(X = 10) = 0.0315$ , determine  $P(2 \leq X \leq 10)$ .

(iii) The average chain length of a polymer produced in a batch reactor is given as 200 units, where chain length itself is known to be a geometric random variable. What fraction of the polymer product is expected to have chains longer than 200 units?

**8.13** The *logarithmic series* random variable possesses the distribution

$$f(x) = \frac{\alpha p^x}{x}; 0 < p < 1; x = 1, 2, \dots, \quad (8.96)$$

First show that the normalizing constant is given by:

$$\alpha = \frac{-1}{\ln(1-p)} \quad (8.97)$$

and then establish the following mathematical characteristics of this random variable and its pdf:

- Mean:  $E(X) = \alpha p / (1 - p)$
- Variance:  $Var(X) = \alpha p(1 - \alpha p)(1 - p)^{-2}$
- Moment generating function:  $M(t) = \ln(1 - pe^t) / \ln(1 - p)$
- Characteristic function:  $\varphi(t) = \ln(1 - pe^{jt}) / \ln(1 - p)$

**8.14** Establish that in the limit as  $k \rightarrow \infty$ , the pdf for the negative binomial  $NBi(k, k/(k + \lambda))$  random variable becomes the pdf for the Poisson  $\mathcal{P}(\lambda)$  random variable.

**8.15** Obtain the recursion formula

$$f(x+1) = \rho(\lambda, x)f(x) \quad (8.98)$$

for the Poisson pdf, showing an explicit expression for  $\rho(\lambda, x)$ . Use this expression to confirm that for all values of  $0 < \lambda < 1$ , the Poisson pdf is always monotonically decreasing. Find the value  $x^*$  for which the pdf attains a maximum for  $\lambda > 1$ . (See comments in Exercise 8.7.)

- 8.16** (i) Obtain the complete pdf,  $f(x)$ , for the binomial random variable with  $n = 10$ ,  $p = 0.05$ , for  $x = 0, 1, \dots, 10$ , and compare it to the corresponding pdf,  $f(x)$ , for a Poisson variable with  $\lambda = 0.5$ .  
(ii) Repeat (i) for  $n = 20$ ,  $p = 0.5$  for the binomial random variable, and  $\lambda = 10$  for the Poisson random variable.

**8.17** Show that if  $X_i, i = 1, 2, \dots, n$ , are  $n$  independent Poisson random variables each with parameter  $\lambda_i$ , then the random variable  $Y$  defined as:

$$Y = \sum_{i=1}^n X_i$$

is also a Poisson random variable, with parameter  $\lambda^* = \sum_{i=1}^n \lambda_i$ .

**8.18** The number of yarn breaks per shift in a commercial fiber spinning machine is a Poisson variable with  $\lambda = 3$ . Determine the probability of not experiencing any yarn break in a particular shift. What is the probability of experiencing more than 3 breaks per shift?

**8.19** The probability of finding a single “fish-eye gel” particle (a solid blemish) on a sq cm patch of a clear adhesive polymer film is 0.0002; the probability of finding more than one is essentially zero. Determine the probability of finding 3 or more such blemishes on a 1 square meter roll of film.

**8.20** For a Poisson  $\mathcal{P}(\lambda)$  random variable, determine  $P(X \leq 2)$  for  $\lambda = 0.5, 1, 2, 3$ . Does the observed behavior of  $P(X \leq 2)$  as  $\lambda$  increases make sense? Explain.

**8.21** The number of eggs laid by a particular bird per mating season is a Poisson random variable  $X$ , with parameter  $\lambda$ . The probability that any such egg successfully develops into a hatchling is  $p$  (the probability that it does not survive is  $(1 - p)$ ). Assuming mutual independence of the development of each egg, if  $Y$  is the random

variable representing the number of surviving hatchlings, establish that its probability distribution function is given by:

$$f(y) = \frac{e^{-p\lambda} (p\lambda)^y}{y!} \quad (8.99)$$

## APPLICATION PROBLEMS

**8.22** (i) A batch of 15 integrated-circuit chips contains 4 of an “irregular” type. If from this batch 2 chips are selected at random, and *without replacement*, find the probability that: (a) both are “irregular”; (b) none is “irregular”; (c) only one of the two is “irregular.”

(ii) If the random variable in problem (i) above was *mistakenly* taken to be a binomial random variable (which it is not), recalculate the three probabilities and compare the corresponding results.

**8.23** A pump manufacturer knows from past records that, in general, the probability of a certain specialty pump working continuously for fewer than 2 years is 0.3; the probability that it will work continuously for 2 to 5 years is 0.5, and the probability that it will work for more than 5 years is 0.2. An order of 8 such pumps has just been sent out to a customer: find the probability that two will work for fewer than 2 years, five will work for 2 to 5 years, and one will work for more than 5 years.

**8.24** The following strategy was adopted in an attempt to determine the size,  $N$ , of the population of an almost extinct population of rare tigers in a remote forest in southeast Asia. At the beginning of the month, 50 tigers were selected from the population, tranquilized, tagged and released; assuming that a month is sufficient time for the tagged sample to become completely integrated with the entire population, at the end of the month, a random sample of  $n = 10$  tigers was selected, two of which were found to have tags.

(i) What does this suggest as a reasonable estimate of  $N$ ? Identify *two* key potential sources of error with this strategy.

(ii) If  $X$  is the random variable representing the total number of tagged tigers found in the sample of  $n$  taken at the end of the month, clearly,  $X$  is a hypergeometric random variable. However, given the comparatively large size we would expect of  $N$  (the unknown tiger population size), it is entirely reasonable to approximate  $X$  as a binomial random variable with a “probability of success” parameter  $p$ . Compute, for this (approximately) binomial random variable, the various probabilities that  $X = 2$  out of the sampled  $n = 10$  when  $p = 0.1$ ,  $p = 0.2$  and  $p = 0.3$ . What does this indicate to you about the “more likely” value of  $p$ , for the tiger population?

(iii) In general, for the binomial random variable  $X$  in (ii) above, given data that  $x = 2$  “successes” were observed in  $n = 10$  trials, show that the probability that  $X = 2$  is maximized if  $p = 0.2$ .

**8.25** The number of contaminant particles (flaws) found on each standard size silicon wafer produced at a certain manufacturing site is a random variable,  $X$ , that a quality control engineer wishes to characterize. A sample of 30 silicon wafers was selected and examined for flaws; the result (the number of flaws found on each wafer)

is displayed in the Table below.

|   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|
| 4 | 1 | 2 | 3 | 2 | 1 | 2 | 4 | 0 | 1 |
| 3 | 0 | 0 | 2 | 3 | 0 | 3 | 2 | 1 | 2 |
| 3 | 4 | 1 | 1 | 2 | 2 | 5 | 3 | 1 | 1 |

- (i) From this data set, obtain an empirical frequency distribution function,  $f_E(x)$ , and compute  $E(X)$ , the expected value of the number of flaws per wafer.
- (ii) Justifying your choice adequately but succinctly, postulate an appropriate theoretical probability model for this random variable. Using the result obtained in (i) above for  $E(X)$ , rounded to the nearest integer, compute from your theoretical model, the probability that  $X = 0, 1, 2, 3, 4, 5$  and 6, and compare these theoretical probabilities to the empirical ones from (i).
- (iii) Wafers with *more than* 2 flaws cannot be sold to customers, resulting in lost revenue; and the manufacturing process ceases to be economically viable if more than 30% of the produced wafers fall into this category. From your theoretical model, determine whether or not the particular process giving rise to this data set is still economically viable.

**8.26** An ensemble of ten identical pumps arranged in parallel is used to supply water to the cooling system of a large, exothermic batch reactor. The reactor (and hence the cooling system) is operated for precisely 8 hrs every day; and the data set shown in the table below is the total number of pumps functioning properly (out of the ten) on any particular 8-hr operating-day, for the entire month of June.

Generate a frequency table for this data set and plot the corresponding histogram. Postulate an appropriate probability model. Obtain a value for the average number of pumps out of 10 that are functioning every day; use this value to obtain an estimate of the model parameters; and from this compute a theoretical pdf. Compare the theoretical pdf with the relative frequency distribution obtained from the data and comment on the adequacy of the model.

| Available Pumps | Day (in June) | Available Pumps | Day (in June) | Available Pumps | Day (in June) |
|-----------------|---------------|-----------------|---------------|-----------------|---------------|
| 9               | 1             | 6               | 11            | 8               | 21            |
| 10              | 2             | 8               | 12            | 9               | 22            |
| 9               | 3             | 9               | 13            | 4               | 23            |
| 8               | 4             | 9               | 14            | 9               | 24            |
| 7               | 5             | 8               | 15            | 9               | 25            |
| 7               | 6             | 9               | 16            | 10              | 26            |
| 9               | 7             | 7               | 17            | 8               | 27            |
| 7               | 8             | 7               | 18            | 5               | 28            |
| 7               | 9             | 7               | 19            | 8               | 29            |
| 8               | 10            | 5               | 20            | 8               | 30            |

**8.27** In a study of the failure of pumps employed in the standby cooling systems of commercial nuclear power plants, Atwood (1986)<sup>2</sup> determined as 0.16 the probability that a pump selected at random in any of these power plants will fail. Consider a

<sup>2</sup>Atwood, C.L., (1986). "The binomial failure rate common cause model," *Technometrics*, 28, 139-148.

system that employs 8 of these pumps but which, for full effectiveness, really requires only 4 to be functioning at any time.

(i) Determine the probability that this particular cooling system will function with full effectiveness.

(ii) If a warning alarm is set to go off when there are five or fewer pumps functioning at any particular time, what is the number of times this alarm is expected to go off in a month of 30 days? State any assumptions you may need to make.

(iii) If the probability of failure increases to 0.2 for each pump, what is the percentage increase in the probability that four or more pumps will fail?

**8.28** The table below contains data from Greenwood and Yule, 1920<sup>3</sup>, showing the frequency of accidents occurring, over a five-week period, to 647 women making high explosives during World War I.

| Number of Accidents | Observed Frequency |
|---------------------|--------------------|
| 0                   | 447                |
| 1                   | 132                |
| 2                   | 42                 |
| 3                   | 21                 |
| 4                   | 3                  |
| 5+                  | 2                  |

(i) For this clearly Poisson-like phenomenon, let  $X$  be the random variable representing the number of accidents. Determine the mean and the variance of  $X$ . What does this indicate about the possibility that this may in fact *not* be a true Poisson random variable?

(ii) Use the value computed for the data average as representative of  $\lambda$  for the Poisson distribution and obtain the theoretical Poisson model prediction of the frequency of occurrences. Compare with the observed frequency.

(iii) Now consider representing this phenomenon as a negative binomial random variable. Determine  $k$  and  $p$  from the computed data average and variance; obtain a theoretical prediction of the frequency of occurrence based on the negative binomial model and compare with the observed frequency.

(iv) To determine, objectively, which probability model provides a better fit to this data set, let  $f_i^o$  represent the observed frequency associated with the  $i^{th}$  group, and let  $\varphi_i$  represent the corresponding theoretical (expected) frequency. For each model, compute the index,

$$C^2 = \sum_{i=1}^m \frac{(f_i^o - \varphi_i)^2}{\varphi_i} \quad (8.100)$$

(For reasons discussed in Chapter 17, it is recommended that group frequencies should not be smaller than 5; as such, the last two groups should be lumped into one group,  $x \geq 4$ .) What do the results of this computation suggest about which model provides a better fit to the data? Explain.

**8.29** Sickle-cell anemia, a serious condition in which the body makes sickle-shaped

---

<sup>3</sup>Greenwood M. and Yule, G. U. (1920) "An enquiry into the nature of frequency distributions representative of multiple happenings with particular reference of multiple attacks of disease or of repeated accidents." *Journal Royal Statistical Society* 83:255-279.

red blood cells, is an inherited disease. People with the disease inherit *two* copies of the sickle cell gene—one from each parent. On the other hand, those who inherit only *one* sickle cell gene from one parent and a normal gene from the other parent have a condition called “sickle cell trait;” such people are sometimes called “carriers” because while they do not have the disease, they nevertheless “carry” one of the genes that cause it and can pass this gene to their children.

Theoretically, if two sickle-cell “carriers” marry, the probability of producing an offspring with the sickle-cell disease is 0.25, while the probability of producing offsprings who are themselves carriers is 0.5; the probability of producing children with a full complement of normal genes is 0.25.

- (i) If a married couple who are both carriers have four children, what is the joint probability distribution of the number of children with sickle cell anemia *and* the number of carriers?
- (ii) From this joint probability distribution, determine the probabilities of having (a) no children with the disease *and* 2 carriers; (b) 1 child with the disease *and* 2 carriers; (c) two children with the disease *and* 2 carriers.
- (iii) On the condition that exactly one of the 4 children is a carrier, determine the probability of having (a) no children with the disease; (b) 1 child with the disease; (c) 2 children with the disease, and (d) 3 children with the disease.

**8.30** Revisit Application Problem 8.29 above and now consider that a married couple, both of whom are carriers, lives in a country where there is no health care coverage, so that each family must cover its own health care costs. The couple, not knowing that that are both carriers, proceed to have 8 children. A child with the sickle-cell disease will periodically experience episodes called “crises” that will require hospitalization and medication for the symptoms (there are no cures yet for the disease). Suppose that it costs the equivalent of US\$2,000 a year in general hospital costs and medication to treat a child with the disease.

- (i) What annual sickle-cell disease-related medical cost can this family expect to incur?
- (ii) If these “crisis” episodes occur infrequently at an average rate of 1.5 per year in this country (3 every two years), and these occurrences are well-modeled as a Poisson-distributed random variable, what is the probability that this family will have to endure a total of 3 crisis episodes in one year? Note that only a child with the disease can have a “crisis” episode. (Hint: See Exercise 8.21.)

**8.31** When a rare respiratory disease with a long incubation period infects a population of people, there is only a probability of 1/3 that an infected patient will show the symptoms within the first month. When five such symptomatic patients showed up in the only hospital in a small town, the astute doctor who treated these patients knew immediately that more will be coming in the next few months as the remaining infected members of the population begin to show symptoms. Assume that all symptomatic patients will eventually come to this one hospital.

- (i) Postulate an appropriate probability model and use it to determine the “most likely” number of infected but not yet symptomatic patients, where  $x^*$  is considered the most likely number if  $P(X = x^*)$  is the highest for all possible values of  $x$ .
- (ii) Because of the nature of the disease, if a *total* of more than 15 people are infected, the small town will have to declare a state of emergency. What is the probability of

this event happening?

**8.32** According to the 1986 *Statistical Abstracts of the United States*, in the five-year period from 1977–1981, failures among banks insured by the Federal Deposit Insurance Corporation (FDIC) averaged approximately 8.5 per year. Specifically, 10 failures were reported in 1980. If FDIC-insured bank failures are considered rare events,

- (i) Postulate an appropriate model for the random variable  $X$  representing the total number of FDIC-insured bank failures per year, and use this model to compute the probability of observing the number of failures reported in 1980.
- (ii) What is the “most likely” number of failures in any one year, if this number,  $x^*$ , is so designated because  $f(x^*)$  is the highest probability of all possible values of  $x$ ? Determine the probability of having more failures in one year than this “most likely” number of failures.
- (iii) An FDIC “quality control inspector” suggested that the occurrence of 13 or more failures in one year should be considered cause for alarm. What is the probability of such an event occurring and why do you think that such an event should truly be a cause for alarm?

**8.33** According to the Welders Association of America, during the decade from 1980–1989, 40% of injuries incurred by its members were to the eye; 22% were to the hand; 20% to the back; and the remaining 18% of injuries were categorized as “others.” Stating whatever assumptions are necessary,

- (i) Determine the probability of recording 4 eye injuries, 3 hand injuries, 2 hand injuries, and 1 injury of the “other” variety.
- (ii) Of 5 total recorded injuries, what is the probability that fewer than 2 are eye injuries?
- (iii) Because eye injuries are the most prevalent, and the most costly to treat, it is desired to reduce their occurrences by investing in eye-safety training programs for the association’s members. What target probability of a single occurrence of an eye injury should the program aim for in order to achieve an overall objective of increasing to approximately 0.9 the probability of observing fewer than 2 eye injuries?

**8.34** On January 28, 1986, on what would be the space shuttle program’s 25<sup>th</sup> mission, the space shuttle *Challenger* exploded. The cause of the accident has since been identified as a failure in the O-ring seal in the solid-fuel booster rocket. A 1983 study commissioned by the Air Force concluded, among other things, that the probability of a catastrophic space shuttle accident due to booster rocket failure is 1/35. Stating whatever assumptions are necessary,

- (i) Determine the probability of attempting 25 missions before the first catastrophe attributable to a booster rocket failure occurs.
- (ii) Determine the probability that the first catastrophe attributable to a booster rocket failure will occur within the first 25 mission attempts (i.e. on or before the 25<sup>th</sup> mission). What does this result imply about the plausibility of the occurrence of this catastrophe at that particular point in time in the history of the space shuttle program?
- (iii) An independent NASA study published in 1985 (just before the accident) claimed that the probability of such a catastrophe happening was 1/60,000. Repeat (ii) above using this value instead of 1/35. In light of the historical fact of the

Jan 28, 1986 indicent, discuss which estimate of the probability the catastrophe's occurrence is more believable?

**8.35** A study in Kalbfleisch *et al.*, 1991<sup>4</sup> reported that the number of warranty claims for one particular system on a particular automobile model within a year of purchase is well-modeled as a Poisson distributed random variable,  $X$ , with an average rate of  $\lambda = 0.75$  claims per car.

(i) Determine the probability that there are two or fewer warranty claims on this specific system for a car selected at random.

(ii) Consider a company that uses the warranty claims on the various systems of the car within the first year of purchase to rate cars for their initial quality. This company wishes to use the Kalbfleisch *et al.* results to set an upper limit,  $x_u$ , on the number of warranty claims whereby a car is declared of "poor initial quality" if the number of claims equals or exceeds this number. Determine the value of  $x_u$  such that, given  $\lambda = 0.75$ , the probability of purchasing a car which, by pure chance alone, will generate more than  $x_u$  warranty claims, is 0.05 or less.

**8.36** In the 1940s, the entomologist S. Corbet catalogued the butterflies of Malaya and obtained the data summarized in the table below. The data table shows  $x$ , a count of the number of species,  $x = 1, 2, \dots, 24$ , and the associated actual number of butterflies caught in light-traps in Malaya that have  $x$  number of species. For example, there were 118 single-species butterflies; 74 two-species butterflies (for a total of 148 of such butterflies) etc, and the last entry indicates that there were 3 categories of 24-species butterflies. Corbet later approached the celebrated R. A. Fisher for assistance in analyzing the data. The result, presented in Fisher *et al.*, 1943<sup>5</sup>, is a record of how the logarithmic series distribution (see Exercise 8.13) was developed as a model for describing species abundance.

Given the characteristics of the logarithmic series distribution in Exercise 8.13, obtain an average for the number of species,  $x$ , and use this to obtain a value for the parameter  $p$  (and hence also  $\alpha$ ). Obtain a predicted frequency  $\hat{\Phi}(x)$  and compare with the values observed in the Corbet data. Comment on the adequacy of this probability model that is now widely used by entomologists for characterizing the distribution of species abundance.

<sup>4</sup>Kalbfleisch, J.D., Lawless, J.F., and Robinson, J.A. (1991). "Methods for the analysis and prediction of warranty claims." *Technometrics*, 33, 273–285.

<sup>5</sup>Fisher, R. A., S. Corbet, and C. B. Williams. (1943). "The relation between the number of species and the number of individuals in a random sample of an animal population." *Journal of Animal Ecology*, 1943: 4258.



# Chapter 9

---

## *Ideal Models of Continuous Random Variables*

|       |   |     |
|-------|---|-----|
| 9.1   | Gamma Family Random Variables .....               | 259 |
| 9.1.1 | The Exponential Random Variable .....             | 259 |
|       | Basic Characteristics and Model Development ..... | 260 |
|       | The Model and Some Remarks .....                  | 261 |
|       | Important Mathematical Characteristics .....      | 261 |
|       | Applications .....                                | 262 |
| 9.1.2 | The Gamma Random Variable .....                   | 264 |
|       | Basic Characteristics and Model Development ..... | 264 |
|       | The Model and Some Remarks .....                  | 265 |
|       | Important Mathematical Characteristics .....      | 266 |
|       | Applications .....                                | 268 |
| 9.1.3 | The Chi-Square Random Variable .....              | 271 |
|       | Basic Characteristics and Model .....             | 271 |
|       | Important Mathematical Characteristics .....      | 271 |
|       | Applications .....                                | 272 |
| 9.1.4 | The Weibull Random Variable .....                 | 272 |
|       | Basic Characteristics and Model Development ..... | 272 |
|       | The Model and Some Remarks .....                  | 273 |
|       | Important Mathematical Characteristics .....      | 274 |
|       | Applications .....                                | 275 |
| 9.1.5 | The Generalized Gamma Model .....                 | 275 |
| 9.1.6 | The Poisson-Gamma Mixture Distribution .....      | 276 |
| 9.2   | Gaussian Family Random Variables .....            | 278 |
| 9.2.1 | The Gaussian (Normal) Random Variable .....       | 279 |
|       | Background and Model Development .....            | 279 |
|       | The Model and Some Remarks .....                  | 287 |
|       | Important Mathematical Characteristics .....      | 288 |
|       | Applications .....                                | 289 |
| 9.2.2 | The Standard Normal Random Variable .....         | 290 |
|       | Important Mathematical Characteristics .....      | 292 |
| 9.2.3 | The Lognormal Random Variable .....               | 292 |
|       | Basic Characteristics and Model Development ..... | 292 |
|       | Important Mathematical Characteristics .....      | 293 |
|       | Applications .....                                | 296 |
| 9.2.4 | The Rayleigh Random Variable .....                | 297 |
|       | Important Mathematical Characteristics .....      | 298 |
|       | Applications .....                                | 300 |
| 9.2.5 | The Generalized Gaussian Model .....              | 300 |
| 9.3   | Ratio Family Random Variables .....               | 300 |
| 9.3.1 | The Beta Random Variable .....                    | 301 |
|       | Basic Characteristics and Model Development ..... | 301 |
|       | The Model and Some Remarks .....                  | 302 |
|       | Important Mathematical Characteristics .....      | 302 |
|       | The Many Shapes of the Beta Distribution .....    | 303 |

|       |  |     |
|-------|--|-----|
| 9.3.2 | Applications .....   | 303 |
|       | Extensions and Special Cases of the Beta Random Variable ..... | 306 |
|       | Generalized Beta Random Variable .....                         | 307 |
|       | Inverted Beta Random Variable .....                            | 307 |
| 9.3.3 | The (Continuous) Uniform Random Variable .....                 | 308 |
|       | Basic Characteristics, Model and Remarks .....                 | 308 |
|       | Important Mathematical Characteristics .....                   | 308 |
|       | Applications .....   | 309 |
| 9.3.4 | Fisher's F Random Variable .....                               | 309 |
|       | Basic Characteristics, Model and Remarks .....                 | 309 |
|       | Important Mathematical Characteristics .....                   | 310 |
|       | Applications .....   | 310 |
| 9.3.5 | Student's t Random Variable .....                              | 311 |
|       | Basic Characteristics, Model and Remarks .....                 | 311 |
|       | Important Mathematical Characteristics .....                   | 312 |
|       | Applications .....   | 314 |
| 9.3.6 | The Cauchy Random Variable .....                               | 314 |
|       | Basic Characteristics, Model and Remarks .....                 | 314 |
|       | Important Mathematical Characteristics .....                   | 315 |
|       | Applications .....   | 316 |
| 9.4   | Summary and Conclusions .....                                  | 316 |
|       | REVIEW QUESTIONS .....   | 317 |
|       | EXERCISES .....  | 323 |
|       | APPLICATION PROBLEMS .....                                     | 329 |

*Facts which at first seem improbable  
 will, even on scant explanation,  
 drop the cloak which has hidden them  
 and stand forth in naked and simple beauty.*

Galileo Galilei (1562–1642)

The presentation of ideal models of randomly varying phenomena that began with discrete random variables in the last chapter concludes in this chapter with the same “first principles” approach applied to continuous random variables. The random variables encountered in this chapter include some of the most celebrated in applied statistics—celebrated because of the central role they play in the theory and practice of statistical inference. Some of these random variables and their pdfs may therefore already be familiar, even to the reader with only rudimentary prior knowledge. But, our objective is not merely to familiarize the reader with continuous random variables and their pdfs; it is to reveal the subterranean roots from which these pdfs sprang, especially the familiar ones. To this end, and as was the case with the discrete random variables of the previous chapter, each model will be derived from the underlying phenomenological characteristics, beginning with the simplest and building up to models for more complex random variables that are themselves functions of several simpler random variables.

The upcoming discussion, even though not exhaustive, is still quite extensive in scope. Because we will be deriving probability models for more than 15 of the most important continuous random variables of practical impor-

tance, to facilitate the discussion and also promote fundamental understanding, these random variables and their pdfs will be presented in “families”—cohort groups that share common structural characteristics. That the starting point of the derivations for the most basic of these families of continuous random variables is a *discrete* random variable may be somewhat surprising at first, but this is merely indicative of the sort of intriguing connections (some obvious, others not) between these random variables—both continuous and discrete. A chart included at the end of the chapter summarizes these connections and places in context how all the random variables discussed in these two chapters are related to one another.

---

## 9.1 Gamma Family Random Variables

Our discussion of continuous random variables and their probability models begins with the “Gamma Family” whose 4 distinct members, from the simplest (in terms of underlying phenomena) to the most complex, are:

- The Exponential random variable,
- The Gamma random variable,
- The Chi-square random variable, and
- The Weibull random variable.

These random variables are grouped together because they share many common structural characteristics, the most basic being non-negativity: they all take values restricted to the positive real line, i.e.  $0 < x < \infty$ . Not surprisingly, they all find application in problems involving intrinsically non-negative entities. Specifically, 3 of the 4 (Exponential, gamma, and Weibull) frequently find application in system reliability and lifetime studies, which involve “waiting times” until some sort of failure. Structurally, these three are much closer together than the fourth, Chi-square, which finds application predominantly in problems involving a *different* class of non-negative variables: mostly squared variables including variances. Its membership in the family is by virtue of being a highly specialized (and somewhat “unusual”) case of the gamma random variable.

### 9.1.1 The Exponential Random Variable

#### Basic Characteristics and Model Development

Let us pick up where we left off in Chapter 8 by considering Poisson events occurring at a constant “intensity”  $\eta$ , (the mean number of “successes” per unit interval size). We wish to consider the random variable,  $X$ , representing the *total interval size* (length of time, spatial length, area, volume, etc) until we observe the *first* occurrence of such Poisson events since the last observation, i.e. the inter-event interval size. (For example, the lifetime of a light bulb—the elapsed time until the filament burns out, or the elapsed time in between the arrival of successive customers at a small town post office counter; the distance between successive flaws on a piece of fibre-optics cable; etc.) The random variable,  $X$ , defined in this manner is an exponential random variable, and its model may be derived from this simple description of its fundamental characteristics as follows.

First, without loss of generality, and simply to help maintain focus on essentials, we shall consider the interval over which the events are happening as *time*, even though it could equally well be length or area or volume. Let us then consider the random variable,  $T$ , the waiting time until we observe the *first* occurrence of these Poisson events.

Now, let  $Y(t)$  be the random variable representing the total *number* of occurrences in the interval  $(0, t)$ , by definition a *discrete* Poisson random variable,  $\mathcal{P}(\eta t)$ . If, as stated above,  $T$  is the time to the first occurrence, then observe that the following probability statement is true:

$$P[Y(t) < 1] = P[T > t] \quad (9.1)$$

because the two mathematical events,  $E_1 = \{y : Y(t) < 1\}$  and  $E_2 = \{t : T > t\}$ , are equivalent. In words: if  $Y(t)$  is not yet 1 (event  $E_1$ ), i.e. if we have not yet observed 1 occurrence, then it is because the current time,  $t$ , is less than the waiting time until the first occurrence (event  $E_2$ ); or equivalently, we have not waited long enough because  $T$ , the waiting time to the first occurrence, is longer than current time  $t$ .

Since  $Y$  is a Poisson random variable with intensity  $\eta$ , we know that:

$$f(y) = \frac{(\eta t)^y e^{-\eta t}}{y!}; y = 0, 1, 2, \dots, \quad (9.2)$$

and since  $P[Y(t) < 1] = P[Y(t) = 0]$ , we obtain from Eq. (9.2) that the expression in Eq (9.1) immediately becomes:

$$\begin{aligned} P[T > t] &= e^{-\eta t}, \text{ or} \\ 1 - F_T(t) &= e^{-\eta t} \end{aligned} \quad (9.3)$$

where  $F_T$  is the cumulative distribution function of the random variable  $T$ , so that:

$$F_T(t) = 1 - e^{-\eta t} \quad (9.4)$$

Upon differentiating once with respect to  $t$ , we obtain the required pdf as

$$f(t) = \eta e^{-\eta t}; 0 < t < \infty, \quad (9.5)$$

the pdf for the exponential random variable,  $T$ , the waiting time until the first occurrence of Poisson events occurring at a constant mean rate,  $\eta$ . This result generalizes straightforwardly from time to spatial intervals, areas, volumes, etc.

### The Model and Some Remarks

In general, the expression

$$f(x) = \eta e^{-\eta x}; 0 < x < \infty \quad (9.6)$$

or, for  $\beta = 1/\eta$ ,

$$f(x) = \frac{1}{\beta} e^{-x/\beta}; 0 < x < \infty \quad (9.7)$$

is the pdf for an exponential random variable,  $\mathcal{E}(\beta)$ .

Recall that we had encountered this same pdf earlier in Chapter 2 as a model for the distribution of residence times in a perfectly mixed continuous stirred tank reactor (CSTR). That model was derived strictly from chemical engineering principles of material balance, with no appeal to probability; this chapter's model arose directly from probabilistic arguments originating from a *discrete* random variable model, with no appeal to physics or engineering. This connection between the highly specialized chemical engineering model and the generic exponential pdf emphasizes the “waiting time” phenomenological attribute of the exponential random variable.

It is also important to note that the geometric distribution discussed in Chapter 8 is a discrete analog of the exponential distribution. The former's phenomenological attribute — the number of discrete trials until the occurrence of a “success” — is so obviously the discrete equivalent of the continuous interval size until the occurrence of a Poisson success that characterizes the latter. Readers familiar with process dynamics and control might also see in the exponential pdf an expression that reminds them of the impulse response of a linear, first order system with steady state gain 1, and time constant,  $\beta$ : in fact the two expressions are identical (after all, the expression in Chapter 2 was obtained as a response of a first order ODE model to an impulse stimulus function). For the purposes of the current discussion, however, the relevant point is that these same readers may now observe that the discrete-time (sampled-data) version of this impulse response bears the same comparison to the expression for the geometric pdf. Thus, the geometric pdf is to the exponential pdf precisely what the discrete time (sampled-data) first order system impulse response function is to the continuous time counterpart.

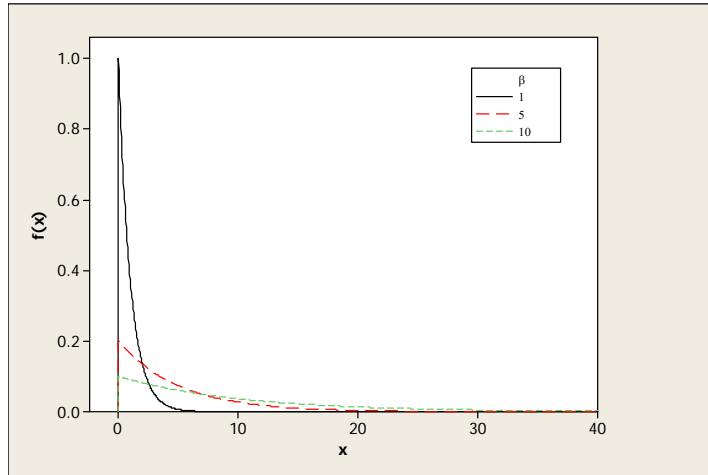


FIGURE 9.1: Exponential pdfs for various values of parameter  $\beta$

### Important Mathematical Characteristics

The following are some important mathematical characteristics of the exponential random variable,  $\mathcal{E}(\beta)$ , and its pdf:

1. **Characteristic parameter:**  $\beta$  (or  $\eta = 1/\beta$ )—the *scale* parameter; it determines how wide the distribution is, with larger values of  $\beta$  corresponding to wider distributions (See Fig 9.1).
2. **Mean:**  $\mu = E(X) = \beta$ .  
(Other measures of central location: Mode = 0; Median =  $\beta \ln 2$ .)
3. **Variance:**  $\sigma^2(X) = \beta^2$
4. **Higher Moments:** Coefficient of Skewness:  $\gamma_3 = 2$ ; Coefficient of Kurtosis:  $\gamma_4 = 9$ , implying that the distribution is positively skewed and sharply peaked. (See Fig 9.1).
5. **Moment generating and Characteristic functions:**

$$M(t) = \frac{1}{(1 - \beta t)} \quad (9.8)$$

$$\varphi(t) = \frac{1}{(1 - j\beta t)} \quad (9.9)$$

6. **Survival function:**  $S(x) = e^{-x/\beta}$   
**Hazard function:**  $h(x) = 1/\beta = \eta$

## Applications

The exponential pdf, not surprisingly, finds application in problems involving waiting times to the occurrence of simple events. As noted earlier, it is most recognizable to chemical engineers as the theoretical residence time distribution function for ideal CSTRs; it also provides a good model for the distribution of time intervals between arrivals at a post office counter, or between phone calls at a customer service center.

Since equipment (or system) reliability and lifetimes can be regarded as waiting times until “failure” of some sort or another, it is also not surprising that the exponential pdf is utilized extensively in reliability and life testing studies. The exponential pdf has been used to model lifetimes of simple devices and of individual components of more complex ones. In this regard, it is important to pay attention to the last characteristic shown above: the constant hazard function,  $h(x)$ . Recall from Chapter 4 that the hazard function allows one to compute the probability of surviving beyond time  $t$ , given survival up to time  $t$ . The constant hazard function indicates that for the exponential random variable, the risk of future failure is independent of current time. The exponential pdf is therefore known as a “memoryless” distribution — the only distribution with this characteristic.

### Example 9.1 WAITING TIME AT SERVICE STATION

The total number of trucks arriving at an all-night service station over the 10:00 pm to 6:00 am night shift is known to be a Poisson random variable with an average hourly arrival rate of 5 trucks/hour. The waiting time between successive arrivals—“idle time” for the service station workers—therefore has the exponential distribution:

$$f(x) = \frac{1}{\beta} e^{-x/\beta}; 0 < x < \infty \quad (9.10)$$

where  $\beta = 1/5$  hours. If the probability is exactly 0.5 that the waiting time between successive arrivals is less than  $\xi$  hours, find the value of  $\xi$ .

#### Solution:

The problem statement translates to:

$$P[X < \xi] = \int_0^\xi 5e^{-5x} dx = 0.5 \quad (9.11)$$

which, upon carrying out the indicated integration, yields:

$$-e^{-5x} \Big|_0^\xi = 0.5 \quad (9.12)$$

or,

$$1 - e^{-5\xi} = 0.5 \quad (9.13)$$

which simplifies to the final desired result:

$$\xi = \frac{-\ln 0.5}{5} = 1.39 \quad (9.14)$$

Note, of course, that by definition,  $\xi$  is the median of the given pdf. The practical implication of this result is that, in half of the arrivals at the service station during this night shift, the waiting (idle) time in between arrivals (on average) will be less than  $\xi = 1.39$  hours; the waiting time will be longer than  $\xi$  for the other half of the arrivals. Such a result can be used in practice in many different ways: for example, the owner of the service station may use this to decide when to hire extra help for the shift, say when the median “idle time” exceeds a predetermined threshold.

### 9.1.2 The Gamma Random Variable

#### Basic Characteristics and Model Development

As a direct generalization of the exponential random variable, consider  $X$ , the random variable defined as the interval size until the  $k^{th}$  occurrence of Poisson events occurring at a constant “intensity”  $\eta$  (e.g., “waiting time” until the occurrence of  $k$  independent events occurring in *time* at a constant mean rate  $\eta$ ).

Again, without loss of generality, and following precisely the same arguments as presented for the exponential random variable, let  $Y(t)$  be the total *number* of occurrences in the interval  $(0, t)$ , but this time, let the random variable,  $T$ , be the time to the  $k^{th}$  occurrence. In this case, the following probability statement is true:

$$P[Y(t) < k] = P[T > t] \quad (9.15)$$

In terms of the implied mathematical events, this statement translates as follows: if  $Y(t)$  is not yet  $k$  (i.e., if we have not yet observed as many as  $k$  total occurrences in the interval  $(0, t)$ , then it is because the current time,  $t$ , is less than the waiting time until the  $k^{th}$  occurrence, (i.e. we have not waited long enough because  $T$ , the waiting time to the  $k^{th}$  occurrence, is longer than current time  $t$ ).

Again, since  $Y$  is a Poisson random variable with intensity  $\eta$ , we know that:

$$P[Y(t) < k] = P[Y(t) \leq (k - 1)] = \sum_{y=0}^{k-1} e^{-\eta t} \frac{(\eta t)^y}{y!} \quad (9.16)$$

The next step in our derivation requires the following result for evaluating the indicated summation:

$$\sum_{y=0}^{k-1} e^{-\eta t} \frac{(\eta t)^y}{y!} = \frac{1}{(k-1)!} \int_{\eta t}^{\infty} e^{-z} z^{k-1} dz \quad (9.17)$$

This result may be obtained (see Exercise 9.6) by first defining:

$$I(k) = \int_a^{\infty} e^{-z} z^{k-1} dz \quad (9.18)$$

and then showing that:

$$I(k) = a^{k-1}e^{-a} + (k-1)I(k-1) \quad (9.19)$$

from where, by recursion, one may then obtain

$$\frac{I(k)}{(k-1)!} = \sum_{y=0}^{k-1} e^{-\eta t} \frac{(\eta t)^y}{y!} \quad (9.20)$$

And now, if  $F_T(t)$  is the cumulative distribution function of the random variable,  $T$ , then the right hand side of Eq (9.15) may be rewritten as

$$P[T > t] = 1 - F_T(t) \quad (9.21)$$

so that the complete expression in Eq (9.15) becomes

$$1 - F_T(t) = \frac{1}{(k-1)!} \int_{\eta t}^{\infty} e^{-z} z^{k-1} dz \quad (9.22)$$

Upon differentiating with respect to  $t$ , using Leibnitz's formula for differentiating under the integral sign, i.e.,

$$\frac{d}{dx} \int_{A(x)}^{B(x)} f(x, r) dr = \int_A^B \frac{\partial f(x, r)}{\partial x} dr + f(x, B) \frac{dB}{dx} - f(x, A) \frac{dA}{dx} \quad (9.23)$$

we obtain:

$$f(t) = \frac{1}{\Gamma(k)} e^{-\eta t} (\eta t)^{k-1} \eta \quad (9.24)$$

where  $\Gamma(k)$  is the *Gamma function* (to be defined shortly), and we have used the fact that for integer  $k$ ,

$$(k-1)! = \Gamma(k) \quad (9.25)$$

The final result,

$$f(t) = \frac{\eta^k}{\Gamma(k)} e^{-\eta t} t^{k-1}; 0 < t < \infty \quad (9.26)$$

is the pdf for the waiting time to the  $k^{th}$  occurrence of independent Poisson events occurring at an average rate  $\eta$ ; it is a particular case of the pdf for a gamma random variable, generalized as follows.

### The Model and Some Remarks

The pdf for the gamma random variable,  $X$ , is given in general by

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} e^{-x/\beta} x^{\alpha-1}; 0 < x < \infty \quad (9.27)$$

the same expression as in Eq (9.26), with  $\beta = 1/\eta$  and replacing  $k$  with a general real number  $\alpha$  not restricted to be an integer.

Some remarks are in order here. First, the random variable name arises from the relationship between the pdf in Eq (9.27) and the *Gamma function* defined by:

$$\Gamma(\alpha) = \int_0^\infty e^{-y} y^{\alpha-1} dy \quad (9.28)$$

If we let  $y = x/\beta$ , we obtain:

$$\Gamma(\alpha) = \frac{1}{\beta^\alpha} \int_0^\infty e^{-x/\beta} x^{\alpha-1} dx \quad (9.29)$$

with the immediate implication that:

$$1 = \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty e^{-x/\beta} x^{\alpha-1} dx \quad (9.30)$$

indicating that the function being integrated on the RHS is a density function. Note also from Eq (9.28), that via integration by parts, one can establish the following well-known recursion property of the gamma function:

$$\Gamma(\alpha + 1) = \alpha \Gamma(\alpha) \quad (9.31)$$

from where it is straightforward to see that if  $\alpha$  is restricted to be a positive integer, then

$$\Gamma(\alpha) = (\alpha - 1)! \quad (9.32)$$

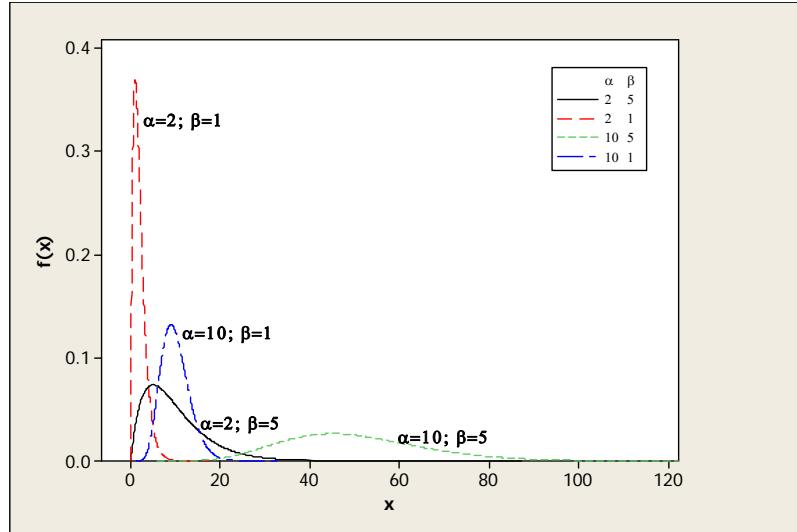
as presented earlier in Eq (9.25).

Second, in the special case when  $\alpha$  is an integer (say  $k$ , as in the preceding derivation), the gamma distribution is known as the Erlang distribution, in honor of the Danish mathematician and engineer, A. K. Erlang (1878-1929) who developed the pdf while working for the Copenhagen Telephone Company. In an attempt to determine how many circuits were needed to provide an acceptable telephone service, Erlang studied the number of telephone calls made to operators at switching stations, in particular, the time between incoming calls. The gamma distribution formally generalizes the Erlang distribution by introducing the real number  $\alpha$  in place of Erlang's integer,  $k$ , and replacing the  $(k - 1)!$  with the gamma function.

Finally, observe that when  $\alpha = 1$ , the resulting pdf is precisely the exponential pdf obtained in the previous subsection. Thus, the exponential random variable is a special case of the gamma random variable, a result that should not be surprising given how the pdf for each random variable was derived.

### Important Mathematical Characteristics

The following are some important mathematical characteristics of the gamma random variable,  $\gamma(\alpha, \beta)$ , and its pdf:



**FIGURE 9.2:** Gamma pdfs for various values of parameter  $\alpha$  and  $\beta$ : Note how with increasing values of  $\alpha$  the shape becomes less skewed, and how the breadth of the distribution increases with increasing values of  $\beta$

1. **Characteristic parameters:**  $\alpha > 0 ; \beta > 0$   
 $\alpha$ , the *shape* parameter, determines overall shape of the distribution (how skewed or symmetric, peaked or flat);  
 $\beta$ , the *scale* parameter, determines how wide the distribution is, with larger values of  $\beta$  corresponding to wider distributions (See Fig 9.2).
2. **Mean:**  $\mu = E(X) = \alpha\beta$ .  
Other measures of central location: Mode =  $\beta(\alpha - 1)$ ;  $\alpha \geq 1$   
Median: No closed-form analytical expression.
3. **Variance:**  $\sigma^2(X) = \alpha\beta^2$
4. **Higher Moments:** Coefficient of Skewness:  $\gamma_3 = 2\alpha^{-1/2}$ ;  
Coefficient of Kurtosis:  $\gamma_4 = 3 + 6/\alpha$ ,  
implying that the distribution is positively skewed but becomes less so with increasing  $\alpha$ , and sharply peaked, approaching the “normal” reference kurtosis value of 3 as  $\alpha \rightarrow \infty$ . (See Fig 9.2).
5. **Moment generating and Characteristic functions:**

$$M(t) = \frac{1}{(1 - \beta t)^\alpha} \quad (9.33)$$

$$\varphi(t) = \frac{1}{(1 - j\beta t)^\alpha} \quad (9.34)$$

**6. Survival function:**

$$S(x) = e^{-x/\beta} \left[ \sum_{i=0}^{\alpha-1} \frac{(x/\beta)^i}{i!} \right] \quad (9.35)$$

valid for the Erlang variable ( $\alpha = \text{integer}$ )

**Hazard function:**

$$h(x) = \frac{(x/\beta)^{\alpha-1}}{\beta \Gamma(\alpha) \sum_{i=0}^{\alpha-1} \frac{(x/\beta)^i}{i!}} \quad (9.36)$$

- 7. Relation to the exponential random variable:** If  $Y_i$  is an exponential random variable with characteristic parameter  $\beta$ , i.e.  $Y_i \sim \mathcal{E}(\beta)$ , then the random variable  $X$  defined as follows:

$$X = \sum_{i=1}^{\alpha} Y_i \quad (9.37)$$

is the gamma random variable  $\gamma(\alpha, \beta)$ . In words: a sum of  $\alpha$  independent and identically distributed exponential random variables is a gamma random variable (more precisely an Erlang random variable because  $\alpha$  will have to be an integer). This result is intuitive from the underlying characteristics of each random variable as presented in the earlier derivations; it is also straightforward to establish using results from Chapter 6 (See Exercise 9.8).

- 8. Reproductive Properties:** The gamma random variable possesses the following useful property: If  $X_i, i = 1, 2, \dots, n$ , are  $n$  independent gamma random variables each with different shape parameters  $\alpha_i$  but a common scale parameter  $\beta$ , i.e.  $X_i \sim \gamma(\alpha_i, \beta)$ , then the random variable  $Y$  defined as:

$$Y = \sum_{i=1}^n X_i \quad (9.38)$$

is also a gamma random variable, with shape parameter  $\alpha^* = \sum_{i=1}^n \alpha_i$  and scale parameter  $\beta$ , i.e.  $Y \sim \gamma(\alpha^*, \beta)$ . Thus, a sum of gamma random variables with identical scale parameters begets another gamma random variable with the same scale parameter, hence the term “reproductive.” (Recall Example 6.6 in Chapter 6.) Furthermore, the random variable  $Z$  defined as

$$Z = c \sum_{i=1}^n X_i \quad (9.39)$$

where  $c$  is a constant, is also a gamma random variable with shape parameter  $\alpha^* = \sum_{i=1}^n \alpha_i$  but with scale parameter  $c\beta$ , i.e.  $Z \sim \gamma(\alpha^*, c\beta)$ . (See Exercise 9.9.)

## Applications

The gamma pdf finds application in problems involving system time-to-failure when system failure occurs as a result of  $\alpha$  independent subsystem failures, each occurring at a constant rate  $1/\beta$ . A standard example is the so-called “standby redundant system” consisting of  $n$  components where system function requires only one component, with the others as backup; when one component fails, another takes over automatically. Complete system failure therefore does not occur until all  $n$  components have failed. For similar reasons, the gamma pdf is also used to study and analyze time between maintenance operations. Because of its flexible shape, the gamma distribution is frequently considered in modeling engineering data of general non-negative phenomena.

As an extension of the application of the exponential distribution in residence time distribution studies in single ideal CSTRs, the gamma distribution may be used for the residence time distribution in several identical CSTRs in series.

The gamma pdf is also used in experimental and theoretical neurobiology, especially in studies involving “action potentials”—the spike trains generated by neurons as a result of nerve-cell activity. These spike trains and the dynamic processes that cause them are random; and it is known that the *distribution* of interspike intervals (ISI) — the elapsed time between the appearance of two consecutive spikes in the spike train — encode information about synaptic mechanisms<sup>1</sup>. Because action potential are generated as a result of a sequence of physiological events, the ISI distribution is often well-modeled by the gamma pdf<sup>2</sup>.

Finally, the gamma distribution has been used recently to model the distribution of distances between DNA replication origins in cells. Fig 9.3, adapted from Chapter 7 of Birtwistle (2008)<sup>3</sup>, shows a  $\gamma(5.05, 8.14)$  distribution fit to data reported in Patel *et al.*, (2006)<sup>4</sup>, on inter-origin distances in the budding yeast *S. cerevisiae*. Note the excellent agreement between the gamma distribution model and the experimental data.

### Example 9.2 REPLACING AUTOMOBILE TIMING BELTS

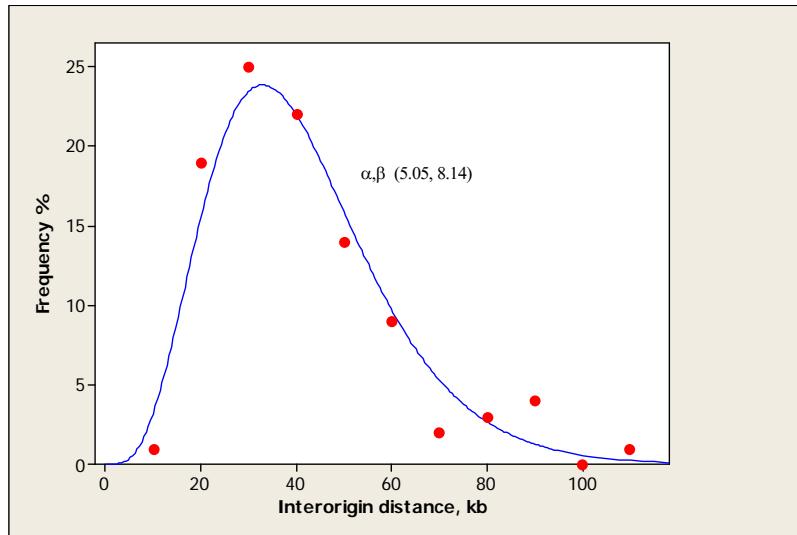
Automobile manufacturers specify in their maintenance manuals the recommended “mileage” at which various components are to be replaced. One such component, the timing belt, must be replaced *before* it breaks, since a broken timing belt renders the automobile entirely inoperative. An extensive experimental study carried out before the official launch of a new model of a certain automobile concluded that  $X$ ,

<sup>1</sup>Braitenberg, 1965: “What can be learned from spike interval histograms about synaptic mechanism?” *J. Theor. Biol.* **8**, 419–425.

<sup>2</sup>H.C. Tuckwell, 1988, *Introduction to Theoretical Neurobiology, Vol 2: Nonlinear and Stochastic Theories*, Chapter 9, Cambridge University Press.

<sup>3</sup>Birtwistle, M. R. (2008). *Modeling and Analysis of the ErbB Signaling Network: From Single Cells to Tumorigenesis*, PhD Dissertation, University of Delaware.

<sup>4</sup>Patel, P. K., Arcangioli, B., Baker, S. P., Bensimon, A. and Rhind, N. (2006). “DNA replication origins fire stochastically in fission yeast.” *Mol Biol Cell* **17**, 308–316.



**FIGURE 9.3:** Gamma distribution fit to data on inter-origin distances in the budding yeast *S. cerevisiae* genome

the lifetimes of the automobile's timing belts (in 10,000 driven miles), is well-modeled by the following pdf:

$$f(x) = \frac{1}{\Gamma(10)} e^{-x} x^9; 0 < x < \infty \quad (9.40)$$

The manufacturer wishes to specify as the recommended mileage (in 10,000 driven miles) at which the timing belt should be replaced, a value  $x^*$ , which is the “most likely” of all possible values of  $X$ , because it maximizes the given  $f(x)$ . Determine  $x^*$  and compare it with the expected value,  $E(X)$  (i.e. the average mileage at failure).

**Solution:**

First, observe that the given pdf is that of a gamma random variable with parameters  $\alpha = 10$ , and  $\beta = 1$ . Next, from the problem statement we see that the desired  $x^*$  is the *mode* of this gamma pdf, in this case:

$$x^* = \beta(\alpha - 1) = 9 \quad (9.41)$$

with the implication that the timing belt should be changed at or before 90,000 driven miles.

For this specific problem,  $E(X) = \alpha\beta = 10$ , indicating that the expected value is 100,000 miles, a value that is longer than the value at the distribution's mode. It appears therefore as if choosing  $x^* = 90,000$  miles as the recommended mileage for the timing belt replacement is a safe and reasonable, if conservative choice. However, since the breakage of a timing belt while the car is in operation on the road is highly

undesirable, recommending replacement at 90,000 miles rather than at the expected average lifetime of 100,000 miles makes sense.

### 9.1.3 The Chi-Square Random Variable

#### Basic Characteristics and Model

An important special case of the gamma random variable arises when  $\beta = 2$  and  $\alpha = r/2$  with  $r$  a positive integer; this is known as the “Chi-square” random variable, with the following model, obtained directly from the gamma pdf:

$$f(x) = \frac{1}{2^{r/2}\Gamma(r/2)} e^{-x/2} x^{\frac{r}{2}-1}; 0 < x < \infty \quad (9.42)$$

This is the pdf for a  $\chi^2(r)$  random variable with  $r$  degrees of freedom. In particular when  $r = 1$ , the resulting random variable,  $\chi^2(1)$ , has the pdf:

$$f(x) = \frac{1}{\sqrt{2}\Gamma(1/2)} e^{-x/2} x^{-1/2}; 0 < x < \infty \quad (9.43)$$

#### Important Mathematical Characteristics

The following mathematical characteristics of the  $\chi^2(r)$  random variable and its pdf derive directly from those of the gamma random variable:

1. **Characteristic parameter:**  $r$ , a positive integer, is the shape parameter, better known as the “degrees of freedom” for reasons discussed later in Part IV in the context of the  $\chi^2(r)$  random variable’s most significant application in statistical inference.
2. **Mean:**  $\mu = r$ ; Mode =  $(r - 2)$ ;  $r \geq 2$
3. **Variance:**  $\sigma^2 = 2r$
4. **Higher Moments:** Coefficient of Skewness:  $\gamma_3 = 2\sqrt{\frac{2}{r}}$ ; Coefficient of Kurtosis:  $\gamma_4 = 3 + 12/r$ .
5. **Moment generating and Characteristic functions:**

$$M(t) = \frac{1}{(1-2t)^{r/2}} \quad (9.44)$$

$$\varphi(t) = \frac{1}{(1-j2t)^{r/2}} \quad (9.45)$$

6. **Reproductive Properties:** The Chi-square random variable inherits from the gamma random variable the following reproductive properties: if  $X_i; i = 1, 2, \dots, k$ , are  $k$  independent  $\chi^2(1)$  random variables, then the random variable  $Y$  defined as  $Y = \sum_{i=1}^k X_i$  has a  $\chi^2(k)$  distribution. Similarly if  $X_i, i = 1, 2, \dots, n$ , are  $n$  independent  $\chi^2(r)$  random variables, then the random variable  $W$  defined as  $W = \sum_{i=1}^n X_i$  has a  $\chi^2(nr)$  distribution. These results find important applications in statistical inference.

Because this random variable is the one family member that is not used in reliability and lifetime studies, we do not provide expressions for the survival and hazard functions.

### Applications

The chi-square random variable is one of a handful of random variables of importance in statistical inference. These applications are discussed more fully in Part IV.

#### 9.1.4 The Weibull Random Variable

To set the stage for the definition of the Weibull random variable, and a derivation of its pdf, let us return briefly to the discussion of the exponential random variable and recall its hazard function,  $h(t) = \eta$ , a constant. According to the definition given in Chapter 4, the corresponding cumulative hazard function (chf),  $H(t)$ , is given by:

$$H(t) = \eta t \quad (9.46)$$

Let us now consider a different situation in which the underlying Poisson random variable's intensity,  $\eta^*$ , is not constant but dependent on the interval length,  $t$ , in such a way that its chf,  $H(t)$ , rather than being a linear function of  $t$  as in Eq (9.46), is given instead by:

$$H(t) = (\eta t)^\zeta \quad (9.47)$$

Thus,

$$\eta^* = h(t) = \frac{d}{dt} H(t) = \zeta \eta (\eta t)^{(\zeta-1)} \quad (9.48)$$

### Basic Characteristics and Model Development

As a generalization of the exponential random variable, consider the random variable,  $T$ , defined as the waiting time to the first occurrence of events occurring this time at a non-constant, time-dependent mean rate  $\eta^*$  given

in Eq (9.48). Then, either by recalling the relationship between the cumulative hazard function and the cumulative distribution function of a random variable, i.e.

$$F_T(t) = 1 - e^{-H(t)} \quad (9.49)$$

or else, following the derivation given above for the exponential random variable, we obtain the cumulative distribution function for this random variable,  $T$  as:

$$F_T(t) = 1 - e^{-(\eta t)^\zeta} \quad (9.50)$$

Upon differentiating once, we obtain:

$$f(t) = \eta \zeta (\eta t)^{\zeta-1} e^{-(\eta t)^\zeta}; 0 < t < \infty \quad (9.51)$$

This is the pdf of a Weibull random variable, named for the Swedish scientist, Waloddi Weibull (1887-1979), who derived and introduced this distribution in a 1939 publication on the analysis of the breaking strength of materials<sup>5</sup>.

It is important now to note that there is something of an “empirical” feel to the Weibull distribution in the sense that if one can conceive of any other reasonable hazard function,  $h(t)$ , with the corresponding chf,  $H(t)$ , such a random variable will have a pdf given by

$$f(t) = h(t)e^{-H(t)} \quad (9.52)$$

There is nothing particularly “phenomenological” about the specific cumulative hazard function,  $H(t)$ , introduced in Eq (9.47) which eventually led to the Weibull distribution; it merely makes the simple linear chf  $H(t) = \eta t$  of the exponential random variable more complex by raising it to a generic power  $\zeta$  — an additional parameter that is to be chosen to fit observations. Unlike the parameter  $\beta$  which has a phenomenological basis, there is no such basis for  $\zeta$ . We shall have cause to return to this point a bit later.

### The Model and Some Remarks

The pdf for the Weibull random variable,  $X$ , is given in general by:

$$f(x) = \eta \zeta (\eta x)^{\zeta-1} e^{-(\eta x)^\zeta}; 0 < x < \infty \quad (9.53)$$

or, for  $\beta = 1/\eta$ ,

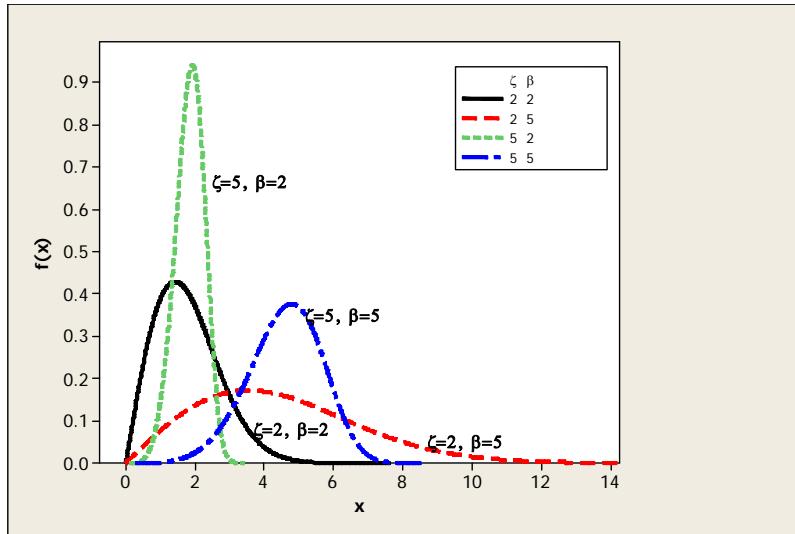
$$f(x) = \frac{\zeta}{\beta} \left( \frac{x}{\beta} \right)^{\zeta-1} e^{-(x/\beta)^\zeta}; 0 < x < \infty \quad (9.54)$$

The model is sometimes written in the following alternative form:

$$f(x) = \beta^* \zeta x^{\zeta-1} e^{-\beta^* x^\zeta} \quad (9.55)$$

---

<sup>5</sup>An earlier independent derivation due to Fisher and Tippet (1928) was unknown to the engineering community until long after Weibull’s work had become widely known and adopted.



**FIGURE 9.4:** Weibull pdfs for various values of parameter  $\zeta$  and  $\beta$ : Note how with increasing values of  $\zeta$  the shape becomes less skewed, and how the breadth of the distribution increases with increasing values of  $\beta$

where

$$\beta^* = \beta^{-\zeta} \quad (9.56)$$

We prefer the form given in Eq (9.54). First,  $\beta$  is the more natural parameter (as we show shortly); second, its role in determining the characteristics of the pdf is distinguishable from that of the second parameter  $\zeta$  in Eq (9.54), whereas  $\beta^*$  in Eq (9.55) is a convolution of the two parameters.

In the special case where  $\zeta = 1$ , the Weibull pdf, not surprisingly, reduces to the exponential pdf. In general, the Weibull random variable,  $W(\zeta, \beta)$ , is related to the exponential random variable  $\mathcal{E}(\beta)$  as follows: if  $Y \sim \mathcal{E}(\beta)$ , then

$$X = Y^{1/\zeta} \quad (9.57)$$

is a  $W(\zeta, \beta)$  random variable; conversely, if  $X \sim W(\zeta, \beta)$  then

$$Y = X^\zeta \quad (9.58)$$

is an  $\mathcal{E}(\beta)$  random variable.

### Important Mathematical Characteristics

The following are some important mathematical characteristics of the Weibull random variable and its pdf:

1. **Characteristic parameters:**  $\zeta > 0$ , and  $\beta > 0$

$\beta$ , as with all the other distribution in this family, is the scale parameter. In this case, it is also known as the “characteristic life” for reasons discussed shortly;

$\zeta$  is the shape parameter. (See Fig 9.4).

2. **Mean:**  $\mu = \beta\Gamma(1 + 1/\zeta)$ ; Mode =  $\beta(1 - 1/\zeta)^{1/\zeta}$ ;  $\zeta \geq 1$ ; Median =  $\beta(\ln 2)^{1/2}$ .
3. **Variance:**  $\sigma^2 = \beta^2 \{\Gamma(1 + 2/\zeta) - [\Gamma(1 + 1/\zeta)]^2\}$
4. **Higher Moments:** Closed form expressions for the coefficients of Skewness and kurtosis are very complex, as are the expressions for the MGF and the Characteristic function.
5. **Survival function:**

$$S(x) = e^{(-x/\beta)^\zeta}; \text{ or } e^{-(\eta x)^\zeta} \quad (9.59)$$

**Hazard function:**

$$h(x) = \frac{\zeta}{\beta} \left(\frac{x}{\beta}\right)^{\zeta-1}, \text{ or } \eta\zeta(\eta x)^{\zeta-1} \quad (9.60)$$

**Cumulative Hazard function:**

$$H(x) = \left(\frac{x}{\beta}\right)^\zeta, \text{ or } (\eta x)^\zeta \quad (9.61)$$

## Applications

The Weibull distribution naturally finds application predominantly in reliability and life-testing studies. It is a very versatile pdf that provides a particularly good fit to time-to-failure data when mean failure rate is time dependent. It is therefore utilized in problems involving lifetimes of complex electronic equipment and of biological organisms, as well as in characterizing failure in mechanical systems. While this pdf is sometimes used to describe the size distribution of particles generated by grinding or crushing operations, it should be clear from the derivation given in this section that such applications are not as “natural” as life-testing applications. When used for particle size characterization, the distribution is sometimes known as the Rosin-Rammler distribution.

An interesting characteristic of the Weibull distribution arises from the following result: when  $x = \beta$ ,

$$P(X \leq \beta) = 1 - e^{-1} = 0.632 \quad (9.62)$$

for *all* values of  $\zeta$ . The parameter  $\beta$  is therefore known as the “characteristic life”; it is the 63.2 percentile of the lifetimes of the phenomenon under study. Readers familiar with process dynamics will recognize the similarity this parameter bears with the time constant,  $\tau$ , of the first order system; in fact,  $\beta$  is to the Weibull random variable what  $\tau$  is to first order dynamic systems.

### 9.1.5 The Generalized Gamma Model

To conclude, we now note that all 4 random variables discussed in this section can be represented as special cases of the generalized gamma random variable,  $X$ , with the following pdf:

$$f(x) = \frac{1}{\beta^{\alpha\zeta}\Gamma(\alpha)} \exp\left[-\left(\frac{x-\delta}{\beta}\right)^{\zeta}\right] \zeta(x-\delta)^{\alpha\zeta-1}; 0 < \delta < x < \infty \quad (9.63)$$

clearly a generalization of Eq (9.27), with the *location* parameter,  $\delta > 0$ ; scale parameter  $\beta$ ; and shape parameters  $\alpha$  and  $\zeta$ . Observe that each pdf discussed in this section may be obtained as special cases of Eq (9.63) as follows:

1. Exponential:  $\alpha = \zeta = 1; \delta = 0$ ;
2. Gamma:  $\zeta = 1; \delta = 0$ ;
3. Chi-squared  $r$ :  $\alpha = r/2, \beta = 2, \zeta = 1; \delta = 0$ ;
4. Weibull:  $\alpha = 1; \delta = 0$ ;

highlighting why these four random variables naturally belong together.

Before leaving the Gamma family of distributions we wish to examine a “mixture distribution” involving the gamma distribution.

### 9.1.6 The Poisson-Gamma Mixture Distribution

We begin with a Poisson random variable,  $X$ , whose now-familiar pdf is:

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}; x = 0, 1, 2, \dots, \quad (9.64)$$

but this time, consider that the parameter  $\lambda$  is *not* constant, but is itself a random variable. This will be the case, for example, if  $X$  represents the number of automobile accidents reported to a company that insures a population of clients for whom the propensity for accidents varies widely. The appropriate model for the entire population of clients will then consist of a mixture of Poisson random variables with different values of  $\lambda$  for the different subgroups within the population. The two most important consequences of this problem definition are as follows:

1. Both  $X$  and  $\lambda$  are random variables, and are characterized by a joint pdf  $f(x, \lambda)$ ;
2. The pdf in Eq (9.64) must now properly be considered as the *conditional*

distribution of  $X$  given  $\lambda$ , i.e.,  $f(x|\lambda)$ ; the expression is a function of  $x$  alone only if the parameter is constant and completely specified (as in Example 8.8 in Chapter 8, where, for the inclusions problem,  $\lambda = 1$ ).

Under these circumstances, what is desired is the unconditional (or marginal) distribution for  $X$ ,  $f(x)$ ; and for this we need a marginal distribution  $f(\lambda)$  for the parameter  $\lambda$ .

Let us now consider the specific case where the parameter is a gamma distributed random variable, i.e.,

$$f(\lambda) = \frac{1}{\beta^\alpha \Gamma(\alpha)} e^{-\lambda/\beta} \lambda^{\alpha-1}; 0 < \lambda < \infty \quad (9.65)$$

In this case, recalling the discussions in Chapter 5, we know that by definition, the joint pdf is obtained as

$$f(x, \lambda) = f(x|\lambda)f(\lambda) \quad (9.66)$$

from where we may now obtain the desired marginal pdf  $f(x)$  by integrating out  $\lambda$ , i.e.,

$$f(x) = \int_0^\infty f(x|\lambda)f(\lambda)d\lambda = \int_0^\infty \left( \frac{\lambda^x e^{-\lambda}}{x!} \right) \left( \frac{1}{\beta^\alpha \Gamma(\alpha)} e^{-\lambda/\beta} \lambda^{\alpha-1} \right) d\lambda \quad (9.67)$$

which is easily rearranged to yield:

$$f(x) = \frac{1}{x! \beta^\alpha \Gamma(\alpha)} \int_0^\infty e^{-\lambda/\beta^*} \lambda^{\alpha^*-1} \quad (9.68)$$

where

$$\frac{1}{\beta^*} = 1 + \frac{1}{\beta} \quad (9.69)$$

$$\alpha^* = x + \alpha \quad (9.70)$$

The reason for such a parameterization is that the integral becomes easy to determine by analogy with the gamma pdf, i.e.,

$$\int_0^\infty e^{-\lambda/\beta^*} \lambda^{\alpha^*-1} = (\beta^*)^{\alpha^*} \Gamma(\alpha^*) \quad (9.71)$$

As a result, Eq (9.68) becomes

$$f(x) = \frac{(\beta^*)^{\alpha^*} \Gamma(\alpha^*)}{x! \beta^\alpha \Gamma(\alpha)} \quad (9.72)$$

which, upon introducing the factorial representation of the gamma function, simplifies to:

$$f(x) = \frac{(x + \alpha - 1)!}{(\alpha - 1)x!} \beta^x (1 + \beta)^{-x-\alpha} \quad (9.73)$$

If we now define

$$1 + \beta = \frac{1}{p}; \Rightarrow \beta = \left( \frac{1-p}{p} \right); \quad (9.74)$$

then Eq (9.73) finally reduces to

$$f(x) = \frac{(x + \alpha - 1)!}{(\alpha - 1)!x!} p^\alpha (1 - p)^x \quad (9.75)$$

immediately recognized as the pdf for a negative binomial  $NBi(\alpha, p)$  random variable. This pdf is formally known as the *Poisson-Gamma mixture* (or compound) distribution, because it is composed from a Poisson random variable whose parameter is gamma distributed; it just happens to be *identical* to the negative binomial distribution.

As stated in Section 8.7, therefore, the appropriate model for a Poisson phenomenon with variable characteristic parameter  $\lambda$  is the negative binomial distribution where, from the results obtained here,  $k = \alpha$  and  $p = 1/(1 + \beta)$ . The relationship between the underlying Poisson model's parameter (which is no longer constant) and the resulting negative binomial model is obtained from the gamma pdf used for  $f(\lambda)$ , i.e.,

$$E(\lambda) = \alpha\beta = \frac{\alpha(1-p)}{p} \quad (9.76)$$

which is precisely the same as the expected value of  $X \sim NBi(\alpha, p)$ .

## 9.2 Gaussian Family Random Variables

The next family of continuous random variables consists of the following 3 members:

- The Gaussian (“Normal”) random variable,
- The Lognormal random variable, and
- The Rayleigh random variable

grouped together, once again, because of shared structural characteristics. This time, however, what is shared is not the domain of the random variables. While the first takes values on the entire real line,  $-\infty < x < \infty$ , the other two take only non-negative values,  $0 < x < \infty$ . What is shared is the functional form of the pdfs themselves.

The Gaussian random variable, the first, and defining member of this family, is unquestionably the most familiar of all random variables, finding broad application in a wide variety of problems, most notably in statistical inference. Unfortunately, by the same token, it also is one of the most “misused”, a point we shall return to later. The second, as its name (lognormal) suggests, is “derivative” of the first in the sense that a logarithmic transformation converts it to the first random variable. It also finds application in statistical inference, especially with such strictly non-negative phenomena as household income, home prices, organism sizes, molecular weight of polymer molecules, particle sizes in powders, crystals, granules and other particulate material, etc — entities whose values can vary over several orders of magnitude. The last variable, perhaps the least familiar, is ideal for representing random deviations of hits from a target *on a plane*. It owes its membership in the family to this very phenomenological characteristic — random fluctuations around a target — that is shared in one form or another by other family members.

Once again, our presentation here centers around derivations of each random variable’s probability model, to emphasize the phenomena underlying each one. A generalized model encompassing all family members is shown at the end to tie everything together.

### 9.2.1 The Gaussian (Normal) Random Variable

#### Background and Model Development

The Gaussian random variable, one of the most versatile and most commonly encountered in physical phenomena, is so named for Johann Carl Friedrich Gauss (1777–1855), the prodigious German mathematician and scientist, whose application of the pdf in his analysis of astronomical data helped popularize the probability model.

However, while Gauss’s name is now forever associated with this random variable and its pdf, the first application on record is actually attributed to Abraham de Moivre (1667–1754) in a 1733 paper, and later to Pierre-Simon de Laplace (1749–1827) who used the distribution for a systematic analysis of measurement errors. The term “normal distribution” became a widely accepted synonym in the late 1800’s because of the popular misconception that the pdf represents a “law of nature” which most, if not all, random phenomena follow. The name, and to some extent, the misconception, survive to this day.

As we now show through a series of derivations, the Gaussian random variable is in fact primarily characterized by an accumulation of a large number of small, additive, independent effects. We present three approaches to the development of the probability model for this random variable: (i) as a limit of the binomial random variable; (ii) from a “first principles” analysis of random motion on a line (also known as the Laplace model of errors); and (iii) the Herschel/Maxwell model of random deviations of hits from a target.

### I: Limiting Case of the Binomial Random Variable

Let us begin by considering a binomial random variable,  $X \sim Bi(n, p)$ , in the limit as the number of trials becomes very large, i.e.  $n \rightarrow \infty$ . Unlike the Poisson random variable case, we will not require that  $p$ , the probability of success, shrink commensurately; rather,  $p$  is to remain fixed, with the implication that the mean number of successes,  $\mu_x = np$ , will also continue to increase with  $n$ , as will the observed number of successes,  $X$ , itself.

This motivates us to define a random variable:

$$Y = X - \mu_x \quad (9.77)$$

the total number of successes *in excess* of the theoretical mean (i.e.  $Y$  represents the *deviation* of  $X$  from the theoretical mean value). Observe that  $Y$  is positive when the observed number of successes exceeds the *mean* number of successes, in which case  $X$  will lie “to the right” of  $\mu_x$  on the real line. Conversely, a negative  $Y$  implies that there are fewer successes than the mean number of successes (or equivalently, that there are more failures than the mean number of failures), so that  $X$  will lie “to the left” of  $\mu_x$ . When the number of successes matches the mean value precisely,  $Y = 0$ . If this “deviation variable” is scaled by the standard deviation  $\sigma_x$ , we obtain the “standardized deviation variable”

$$Z = \frac{X - \mu_x}{\sigma_x} \quad (9.78)$$

It is important to note that even though  $X$ ,  $\mu_x$  and  $\sigma_x$  can each potentially increase indefinitely as  $n \rightarrow \infty$ ,  $Z$  on the other hand is “well-behaved:” regardless of  $n$ , its expected value  $E(Z) = \mu_z = 0$  and its variance  $\sigma_z^2 = 1$  (see Exercise 9.15).

We are now interested first in obtaining a pdf,  $f(z)$ , for  $Z$  the standardized deviation of the binomial random variable from its theoretical mean value, in the limit of a large number of trials; the desired  $f(x)$  for  $X$  will then be recovered from  $f(z)$  and the transformation in Eq (9.78).

Out of a variety of ways of obtaining the pdf,  $f(z)$ , we opt for the method of characteristic functions. Recall that the characteristic function for the binomial random variable,  $X$ , is

$$\varphi_x(t) = \{pe^{jt} + q\}^n \quad (9.79)$$

and from the properties of the MGF and CF given in Chapter 4, we obtain from Eq (9.78) that

$$\varphi_z(t) = e^{\left(\frac{-jtnp}{\sigma_x}\right)} \left\{ pe^{jt/\sigma_x} + q \right\}^n \quad (9.80)$$

whereupon taking natural logarithms yields

$$\ln \varphi_z(t) = \frac{-jtnp}{\sigma_x} + n \ln \left\{ 1 + p \left( e^{jt/\sigma_x} - 1 \right) \right\} \quad (9.81)$$

having introduced  $(1 - p)$  for  $q$ , the complementary probability of failure. A Taylor series expansion of the exponential term yields:

$$\ln \varphi_z(t) = \frac{-jtnp}{\sigma_x} + n \ln \left\{ 1 + p \left[ \left( \frac{jt}{\sigma_x} \right) - \frac{1}{2} \left( \frac{t}{\sigma_x} \right)^2 + \mathcal{O}(\sigma_x^{-3}) \right] \right\} \quad (9.82)$$

where  $\mathcal{O}(\sigma_x^{-3})$  is a term that goes to zero faster than  $\sigma_x^{-3}$  as  $n \rightarrow \infty$  (recall that for the binomial random variable,  $\sigma_x = \sqrt{npq}$ , so that  $\sigma_x^{-3} \rightarrow 0$  as  $n \rightarrow \infty$ ). If we now invoke the result that

$$\ln(1 + \alpha) = \alpha - \frac{\alpha^2}{2} + \frac{\alpha^3}{3} - \dots \quad (9.83)$$

with

$$\alpha = p \left[ \left( \frac{jt}{\sigma_x} \right) - \frac{1}{2} \left( \frac{t}{\sigma_x} \right)^2 + \mathcal{O}(\sigma_x^{-3}) \right] \quad (9.84)$$

so that,

$$\alpha^2 = \frac{-p^2 t^2}{\sigma_x^2} + \mathcal{O}(\sigma_x^{-3}) \quad (9.85)$$

then the second term on the RHS in Eq (9.82) reduces to:

$$n \left\{ \frac{pj}{\sigma_x} - \frac{1}{2} \frac{t^2}{\sigma_x^2} (p - p^2) + \mathcal{O}(\sigma_x^{-3}) \right\}$$

From here, the entire Eq (9.82) becomes:

$$\ln \varphi_z(t) = \frac{-jtnp}{\sigma_x} + \frac{jtnp}{\sigma_x} - \frac{1}{2} t^2 \frac{np(1-p)}{\sigma_x^2} + \mathcal{O}(\sigma_x^{-3}) \quad (9.86)$$

which, in the limit as  $n \rightarrow \infty$ , simplifies to

$$\ln \varphi_z(t) = -\frac{1}{2} t^2 \quad (9.87)$$

so that:

$$\varphi_z(t) = e^{-t^2/2} \quad (9.88)$$

To obtain the corresponding  $f(z)$ , we may now simply consult compilations of characteristic functions, or else from the definition of characteristic function and pdf pairs given in Chapter 4, obtain  $f(z)$  from the integral:

$$f(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-jtz} e^{-t^2/2} dt \quad (9.89)$$

and upon carrying out the indicated integration, we obtain the required pdf:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad (9.90)$$

as the pdf of the standardized deviation variable,  $Z$  defined in Eq (9.78). From the relationship between  $Z$  and  $X$ , it is now a straightforward matter to recover the pdf for  $X$  as (see Exercise 9.15):

$$f(x) = \frac{1}{\sigma_x \sqrt{2\pi}} e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2}} \quad (9.91)$$

This is the pdf of a Gaussian random variable with mean value  $\mu_x$  and standard deviation  $\sigma_x$  inherited from the original binomial random variable.

## II: First-Principles (Random Motion in a Line)

Consider a particle moving randomly along the  $x$ -axis with motion governed by the following rules:

1. Each move involves taking a single step of fixed length,  $\Delta x$ , once every time interval  $\Delta t$ ;
2. The step can be to the right (with probability  $p$ ), or to the left (with probability  $q = 1 - p$ ): for simplicity, this presentation will consider equal probabilities,  $p = q = 1/2$ ; the more general derivation is a bit more complicated but the final result is the same.

We are interested in the probability of finding a particle  $m$  steps to the right of the starting point after making  $n$  independent moves, in the limit as  $n \rightarrow \infty$ .

Before engaging in the model derivation, the following are some important points about the integer  $m$  that will be useful later:

1.  $m$  can be negative or positive and is restricted to lie between  $-n$  and  $n$ ;
2. If  $k > 0$  is the total number of steps taken to the right (so that the total number of steps taken to the left is  $(n - k)$ ), for the particle to reach a point  $m$  steps to the right implies that:

$$m = k - (n - k) = (2k - n)$$

so that:

$$k = \frac{1}{2}(n + m)$$

3.  $m = 2k - n$  must be even when  $n$  is even, and odd when  $n$  is odd; therefore
4.  $m$  ranges from  $-n$  to  $n$  in steps of 2. For example, if  $n = 3$  (3 total steps taken) then  $m$  can take only the values  $-3, -1, 1$ , or  $3$ ; and for  $n = 4$ ,  $m$  can only be  $-4, -2, 0, 2$  or  $4$ .

Now, define as  $P(x, t)$ , the probability that a particle starting at the origin at time  $t = 0$  arrives at a point  $x$  at time  $t$ , where,

$$x = m\Delta x; t = n\Delta t \quad (9.92)$$

It is then true that:

$$P(x, t + \Delta t) = \frac{1}{2}P(x - \Delta x, t) + \frac{1}{2}P(x + \Delta x, t) \quad (9.93)$$

To reach point  $x$  at time  $(t + \Delta t)$ , then at time  $t$ , one of two events must happen: (i) the particle must reach point  $(x - \Delta x)$  (with probability  $P(x - \Delta x, t)$ ) and then take a step to the right (with probability 1/2); or (ii) the particle must reach point  $(x + \Delta x)$  (with probability  $P(x + \Delta x, t)$ ), and then take a step to the left (with probability 1/2). This is what is represented by Eq (9.93); in the limit as  $n \rightarrow \infty$ , it provides an expression for the pdf we wish to derive. The associated initial conditions are:

$$\begin{aligned} P(0, 0) &= 1 \\ P(x, 0) &= 0; x \neq 0 \end{aligned} \quad (9.94)$$

indicating that, since we began at the origin, the probability of finding this particle at the origin, at  $t = 0$ , is 1; and also that the probability of finding the particle, at this same time  $t = 0$ , at any other point,  $x$  that is *not* the origin, is 0.

Now, as  $n \rightarrow \infty$ , for  $t$  to remain fixed,  $\Delta t$  must tend to zero; similarly, as  $n \rightarrow \infty$ , so must  $m \rightarrow \infty$ , and for  $x$  to remain fixed,  $\Delta x$  must tend to zero as well. However, by definition,  $m < n$  so that as both become large,  $n \rightarrow \infty$  faster than  $m$  so that  $m/n \rightarrow 0$ ; i.e.

$$\frac{m}{n} = \frac{x}{\Delta x} \frac{\Delta t}{t} = \frac{x}{t} \frac{\Delta t}{\Delta x} \rightarrow 0 \quad (9.95)$$

implying that  $\Delta x/\Delta t \rightarrow \infty$  but in such a way that

$$\Delta x \left( \frac{\Delta x}{\Delta t} \right) = \text{fixed} \quad (9.96)$$

The importance of this point will soon become clear.

Now, subtracting  $P(x, t)$  from both sides of Eq (9.93) yields:

$$P(x, t + \Delta t) - P(x, t) = \frac{1}{2}P(x - \Delta x, t) - \frac{1}{2}P(x, t) + \frac{1}{2}P(x + \Delta x, t) - \frac{1}{2}P(x, t) \quad (9.97)$$

Upon multiplying the left hand side by  $(\Delta t/\Delta t)$  and the right hand side by  $(\Delta x^2/\Delta x^2)$  and rearranging appropriately, we obtain:

$$\begin{aligned} \text{LHS} &= \left[ \frac{P(x, t + \Delta t) - P(x, t)}{\Delta t} \right] \Delta t \\ \text{RHS} &= \frac{1}{2}(\Delta x)^2 \left\{ \frac{1}{\Delta x} \left[ \frac{P(x + \Delta x, t) - P(x, t)}{\Delta x} - \frac{P(x, t) - P(x - \Delta x, t)}{\Delta x} \right] \right\} \end{aligned} \quad (9.98)$$

And now, recalling Eq (9.96) and defining

$$\lim_{\substack{\Delta x \rightarrow 0 \\ \Delta t \rightarrow 0}} \frac{(\Delta x^2)}{2\Delta t} = \mathcal{D} \neq 0 \quad (9.99)$$

(where  $\mathcal{D}$  is a fixed constant), then upon taking limits as  $\Delta x \rightarrow 0$  and  $\Delta t \rightarrow 0$  above, we obtain:

$$\frac{\partial P(x, t)}{\partial t} = \mathcal{D} \frac{\partial^2 P(x, t)}{\partial x^2} \quad (9.100)$$

an equation that may be recognizable to readers familiar with the physical phenomenon of diffusion.

Before proceeding to solve this equation for  $P(x, t)$  however, we need to recognize that, strictly speaking, as the number of points becomes infinite (as  $n, m \rightarrow \infty$  and  $\Delta x \rightarrow 0$  and  $\Delta t \rightarrow 0$ ), the probability of finding a particle at any arbitrary point  $x$  tends to zero. To “regularize” this function, let us recall the nature of  $m$  (which indicates that the particle can occupy only *every other point* on the  $x$ -axis), and introduce the function

$$f(x, t) = \frac{P(x, t)}{2\Delta x} \quad (9.101)$$

As a result, the probability of finding a particle, at time  $t$ , on or between two points  $a_1 = i\Delta x$  and  $a_2 = k\Delta x$  is properly given by:

$$P(a_1 < x < a_2; t) = \sum_{m=i}^k f(m\Delta x, t) 2\Delta x \quad (9.102)$$

(keeping in mind that the sum is for every other value of  $m$  from  $i$  to  $k$ , with both indices even if  $n$  is even, and odd if  $n$  is odd). Eq (9.100) now becomes

$$\frac{\partial f(x, t)}{\partial t} = \mathcal{D} \frac{\partial^2 f(x, t)}{\partial x^2} \quad (9.103)$$

which is more mathematically precise for determining probabilities in the limit as the original function becomes continuous, even though it looks almost exactly like Eq (9.100). Observe that in the limit as  $\Delta x \rightarrow 0$ , Eq (9.102) becomes:

$$P(a_1 < x < a_2; t) = \int_{a_1}^{a_2} f(x, t) dx \quad (9.104)$$

so that, as the reader may perhaps have suspected all along,  $f(x, t)$  is the required (time dependent) probability density function for the random phenomenon in question — random motion on a line.

We are now in a position to solve Eq (9.103), but the initial conditions for  $P(x, t)$  in Eq (9.94) must now be modified appropriately for  $f(x, t)$  as follows:

$$\begin{aligned} \int_{-\infty}^{\infty} f(x, t) dx &= 1 \\ \lim_{t \rightarrow 0} f(x, t) &= 0 \end{aligned} \quad (9.105)$$

The first implies that, at any point in time, the particle will, with certainty, be located somewhere on the  $x$ -axis; the second is the continuous equivalent of

the condition on  $P(x, 0)$  in Eq (9.94), a requirement for the particle starting at the origin, at  $t = 0$ .

Readers familiar with partial differential equations in general, or the one shown in Eq (9.103) in particular, will be able to confirm that the solution, subject to the conditions in Eq (9.105), is:

$$f(x, t) = \frac{1}{\sqrt{4\pi D t}} e^{\frac{-x^2}{4Dt}} \quad (9.106)$$

If we now return to the definition of the parameter  $D$  in Eq (22.58), we see that:

$$2Dt = \Delta x^2 \frac{t}{\Delta t} = n(\Delta x^2) \quad (9.107)$$

which is the total sum of squared displacements (fluctuations) to the right and to the left in time  $t$ . If we represent this measure of the vigor of the dispersion as  $b^2$ , we obtain the expression,

$$f(x) = \frac{1}{b\sqrt{2\pi}} e^{\frac{-x^2}{2b^2}} \quad (9.108)$$

where the time argument  $t$  has been suppressed because it is no longer explicit (having been subsumed into the dispersion parameter  $b$ ). Finally, for an arbitrary starting point  $a \neq 0$  the required pdf becomes:

$$f(x) = \frac{1}{b\sqrt{2\pi}} e^{\frac{-(x-a)^2}{2b^2}} \quad (9.109)$$

which is to be compared with the expression obtained earlier in Eq (9.91).

### III: Herschel/Maxwell Model

Consider an experiment in which small pellets (or grains of sand) are dropped unto a plane from above the point labeled O as illustrated in Fig 9.5. We are interested in the probability of finding one of the pellets on the shaded element of area  $\Delta A$  under the following conditions:

1. *Symmetry:* There is no systematic deviation of pellets from the center point, O; i.e. deviations are purely random so that for a given  $r$ , all angular positions,  $\theta$ , are possible. *As a result, the probability of finding a pellet in a small area at a distance  $r$  from the center is the same for all such areas at the same distance.*
2. *Independence:* The probability of finding a pellet between the points  $x$  and  $x + \Delta x$  in the horizontal coordinate is completely independent of the pellet position in the vertical,  $y$  coordinate.

From point 1 above, the pdf we seek must be a function of  $r$  alone, say  $g(r)$ ,

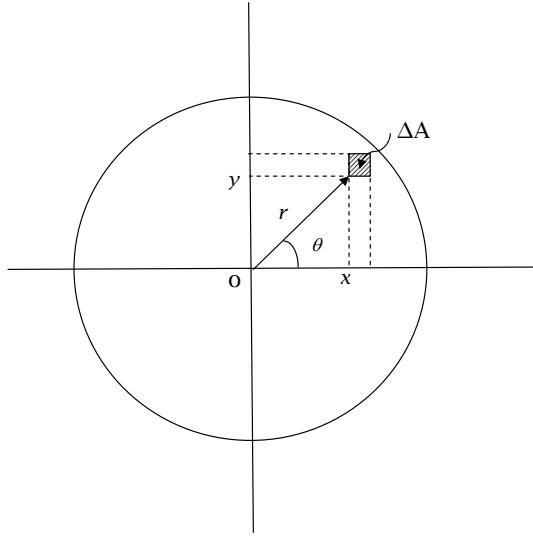


FIGURE 9.5: The Herschel-Maxwell 2-dimensional plane

where, of course, from standard analytical geometry, the following relationships hold:

$$\begin{aligned} x &= r \cos \theta; \\ y &= r \sin \theta; \\ r^2 &= x^2 + y^2. \end{aligned} \quad (9.110)$$

From point 2, the joint pdf satisfies the condition that

$$f(x, y) = f(x)f(y). \quad (9.111)$$

As a result, the probability of finding a pellet in an area of size  $\Delta A$  is

$$g(r)\Delta A = f(x)f(y)\Delta A \quad (9.112)$$

which, upon taking logs and using Eqs (9.110), yields

$$\ln g(r) = \ln f(r \cos \theta) + \ln f(r \sin \theta) \quad (9.113)$$

Differentiating with respect to  $\theta$  results in:

$$0 = -r \sin \theta \frac{f'(r \cos \theta)}{f(r \cos \theta)} + r \cos \theta \frac{f'(r \sin \theta)}{f(r \sin \theta)} \quad (9.114)$$

where the 0 on the LHS arises because  $r$  is independent of  $\theta$ . If we now return to cartesian coordinates, we obtain

$$\frac{f'(x)}{xf(x)} = \frac{f'(y)}{yf(y)} \quad (9.115)$$

It is now clear that  $x$  and  $y$  are entirely independent since the LHS is a function of  $x$  alone and the RHS is a function of  $y$  alone; furthermore, these two will then be equal only if they are both equal to a constant, say  $c_1$ , i.e.

$$\frac{f'(x)}{f(x)} = c_1 x \quad (9.116)$$

Integrating and rearranging leads to

$$\begin{aligned} \ln f(x) &= \frac{1}{2}c_1 x^2 + c_2; \text{ or} \\ f(x) &= ke^{\frac{1}{2}c_1 x^2} \end{aligned} \quad (9.117)$$

where the new constant  $k = \exp(c_2)$ .

We now need to determine the integration constants. Because  $f(x)$  must be a valid pdf, it must remain finite as  $x \rightarrow \infty$ ; this implies immediately that  $c_1$  must be negative, say  $c_1 = -1/b^2$ , with the result that:

$$f(x) = ke^{-\frac{1}{2}\frac{x^2}{b^2}} \quad (9.118)$$

Simultaneously, because of Eq (9.115),

$$f(y) = ke^{-\frac{1}{2}\frac{y^2}{b^2}} \quad (9.119)$$

so that

$$g(r) = f(x)f(y) = k^2 e^{-\frac{1}{2}\frac{r^2}{b^2}} \quad (9.120)$$

In general, if the point O is not the origin but some other arbitrary point  $(a_x, a_y)$  in the plane, then Eq (9.118) becomes:

$$f(x) = ke^{-\frac{(x-a_x)^2}{2b^2}} \quad (9.121)$$

And now, since  $\int f(x)dx = 1$ , and it can be shown that

$$\int_{-\infty}^{\infty} e^{-\frac{(x-a_x)^2}{2b^2}} dx = b\sqrt{2\pi}, \quad (9.122)$$

then it follows that  $k = 1/b\sqrt{2\pi}$  so that the required pdf is given by:

$$f(x) = \frac{1}{b\sqrt{2\pi}} e^{-\frac{(x-a_x)^2}{2b^2}} \quad (9.123)$$

exactly as we had obtained previously.

### The Model and Some Remarks

The pdf for a Gaussian random variable is

$$f(x) = \frac{1}{b\sqrt{2\pi}} \exp\left\{\frac{-(x-a)^2}{2b^2}\right\}; \infty < x < \infty; b > 0 \quad (9.124)$$

a model characterized by two parameters,  $a$ , the location parameter, and  $b$ , the scale parameter. Now, because,  $E(X) = \mu = a$ ; and  $Var(X) = \sigma^2 = b^2$ , the more widely encountered form of the pdf is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\} \quad (9.125)$$

which co-opts the universal symbols for mean and standard deviation for this particular random variable's mean and standard deviation. A variable with this pdf is said to possess a  $N(\mu, \sigma^2)$  distribution, "normal, with mean  $\mu$ , and variance  $\sigma^2$ ."

Some remarks are now in order. The derivations shown above indicate that the random phenomenon underlying the Gaussian random variable is characterized by:

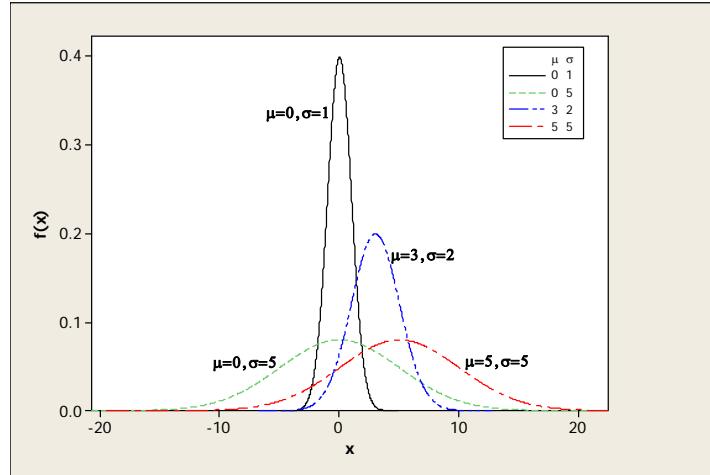
1. Observations composed of the net effect of many small, additive, perturbations, some negative, some positive; or
2. Random, symmetric, independent deviations from a "target" or true value.

Thus, such diverse phenomena as measurement errors, heights and weights in human populations, velocities of gas molecules in a container, and even test scores, all tend to follow the Gaussian distribution. However, it is important to note that the Gaussian distribution is *not a law of nature*, contrary to a popular misconception. The synonym "Normal" used to describe this pdf tends to predispose many to assume that just about any random variable has a Gaussian distribution. Of course, this is clearly not true. (Recall that we have already discussed many random variables that do not follow the Gaussian distribution.)

This misconception of "near-universality" of the Gaussian random variable is also fueled by a property of averages of random variables discussed later in Part 4 when we examine the Central Limit Theorem. For now, we caution the reader to be careful how the  $N(\mu, \sigma^2)$  distribution is assumed for a random variable: if the underlying characteristics are not reasonably close to the ones discussed above, it is unlikely that the Gaussian distribution is appropriate.

### Important Mathematical Characteristics

The following are some key mathematical characteristics of the  $N(\mu, \sigma^2)$  random variable and its pdf:



**FIGURE 9.6:** Gaussian pdfs for various values of parameter  $\mu$  and  $\sigma$ : Note the symmetric shapes, how the center of the distribution is determined by  $\mu$ , and how the shape becomes broader with increasing values of  $\sigma$

1. **Characteristic parameters:**  $\mu$ , the location parameter, is also the mean value;  $\sigma$ , the scale parameter, is also the standard deviation;
2. **Mean:**  $E(X) = \mu$ ; Mode =  $\mu$  = Median
3. **Variance:**  $Var(X) = \sigma^2$
4. **Higher Moments:** Coefficient of Skewness:  $\gamma_3 = 0$ ; Coefficient of Kurtosis:  $\gamma_4 = 3$ . This value of 3 is the standard reference for kurtosis alluded to in Chapter 4. Recall that distributions for which  $\gamma_4 < 3$  are said to be *platykurtic* (mildly peaked) while those for which  $\gamma_4 > 3$  are said to be *leptokurtic* (sharply peaked).
5. **Moment generating and Characteristic functions:**

$$M(t) = \exp\{\mu t + \frac{1}{2}\sigma^2 t^2\} \quad (9.126)$$

$$\varphi(t) = \exp\{j\mu t - \frac{1}{2}\sigma^2 t^2\} \quad (9.127)$$

6. The function is perfectly symmetric about  $\mu$ ; Also, at  $x = \mu$ ,  $f'(x) = 0$  and  $f''(x) < 0$  establishing  $x = \mu$  also as the mode; finally,  $f''(x) = 0$  at  $x = \pm\sigma$ . See Fig 9.6.

### Applications

The Gaussian random variable plays an important role in statistical inference where its applications are many and varied. While these applications are discussed more fully in Part IV, we note here that they all involve computing probabilities using  $f(x)$ , or else, given specified tail area probabilities, using  $f(x)$  in reverse to find the corresponding  $x$  values. Nowadays, the tasks of carrying out such computations have almost completely been delegated to computer programs; traditionally, practical applications required the use of pre-computed tables of normal cumulative probability values. Because it is impossible (and unrealistic) to generate tables for all conceivable values of  $\mu$  and  $\sigma$ , the traditional normal probability tables are based on the standard normal random variable,  $Z$ , which, as we now discuss, makes it possible to apply these tables for all possible values of  $\mu$  and  $\sigma$ .

#### 9.2.2 The Standard Normal Random Variable

If the random variable  $X$  possesses a  $N(\mu, \sigma^2)$  distribution, then the random variable defined as:

$$Z = \frac{X - \mu}{\sigma} \quad (9.128)$$

has the pdf

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\} \quad (9.129)$$

$Z$  is called a standard normal random variable; its mean value is 0, its standard deviation is 1, i.e. it has a  $N(0, 1)$  distribution. A special case of the general Gaussian random variable, its traditional utility derives from the fact that for any general Gaussian random variable  $X \sim N(\mu, \sigma^2)$ ,

$$P(a_1 < X < a_2) = P\left(\frac{a_1 - \mu}{\sigma} < Z < \frac{a_2 - \mu}{\sigma}\right) \quad (9.130)$$

so that tables of  $N(0, 1)$  probability values for various values of  $z$  can be used to compute probabilities for any and all general  $N(\mu, \sigma^2)$  random variables.

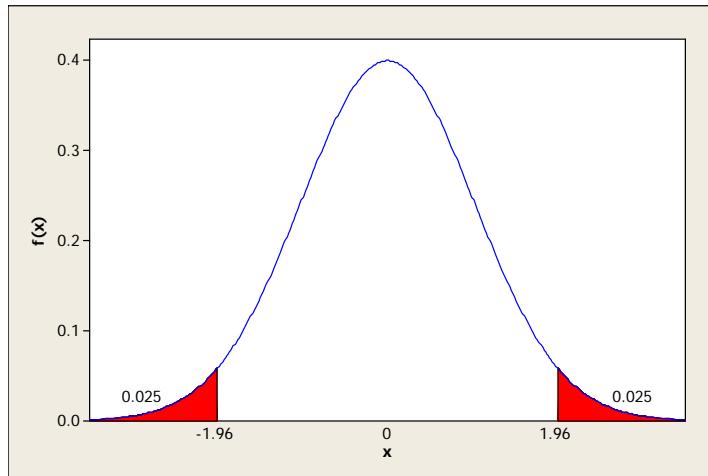
The *z-score* of any particular value  $x_i$  of the general Gaussian random variable  $X \sim N(\mu, \sigma^2)$  is defined as

$$z_i = \frac{x_i - \mu}{\sigma} \quad (9.131)$$

Probabilities such as those shown in Eq (9.130) are therefore determined on the basis of the “z-scores” of the indicated values  $a_1$  and  $a_2$ .

Furthermore, because the distribution is symmetric about  $x = 0$ , it is true that:

$$F_Z(-a) = 1 - F_Z(a) \quad (9.132)$$



**FIGURE 9.7:** Symmetric tail area probabilities for the standard normal random variable with  $z = \pm 1.96$  and  $F_Z(-1.96) = 0.025 = 1 - F_Z(1.96)$

where  $F_Z(a)$  is the cumulative probability defined in the usual manner as:

$$F_Z(a) = \int_{-\infty}^a f(z) dz \quad (9.133)$$

Figure 9.7 shows this for the specific case where  $a = 1.96$  for which the tail areas are each 0.025.

This result has the implication that tables of tail area probabilities need only be made available for positive values of  $Z$ . The following example illustrates this point.

#### Example 9.3 POST-SECONDARY EXAMINATION TEST SCORES

A collection of all the test scores for a standardized, post-secondary examination administered in the 1970's across countries along the West African coast, is well-modeled as a random variable  $X \sim N(\mu, \sigma^2)$  with  $\mu = 270$  and  $\sigma = 26$ . If a score of 300 or higher is required for a "pass-with-distinction" grade, and a score between 260 and 300 is required for a "merit-pass" grade, what percentage of students will receive the "distinction" grade and what percentage will receive the "merit" grade?

#### Solution:

The problem requires computing the following probabilities:  $P(X \geq 300)$  and  $P(260 < X < 300)$ .

$$\begin{aligned} P(X \geq 300) &= 1 - P(X < 300) = 1 - P\left[Z < \left(\frac{300 - 270}{26}\right)\right] \\ &= 1 - F_Z(1.154) \end{aligned} \quad (9.134)$$

indicating that the  $z$ -score for  $x = 300$  is 1.154. From tables of cumulative probabilities for the standard normal random variable, we obtain  $F_Z(1.154) = 0.875$  so that the required probability is given by

$$P(X \geq 300) = 0.125 \quad (9.135)$$

implying that 12.5% of the students will receive the “distinction” grade.

The second probability is obtained as:

$$\begin{aligned} P(260 < X < 300) &= P\left[\left(\frac{260 - 270}{26}\right) < Z < \left(\frac{300 - 270}{26}\right)\right] \\ &= F_Z\left(\frac{30}{26}\right) - F_Z\left(\frac{-10}{26}\right) \end{aligned} \quad (9.136)$$

And now, by symmetry,  $F(-10/26) = 1 - F(10/26)$  so that, from the cumulative probability tables, we now obtain:

$$P(260 < X < 300) = 0.875 - (1 - 0.650) = 0.525 \quad (9.137)$$

with the implication that 52.5% of the students will receive the “merit” grade.

Of course, with the availability of such computer programs as MINITAB, it is possible to obtain the required probabilities directly without recourse to the  $Z$  probability tables. In this case, one simply obtains  $F_X(300) = 0.875$  and  $F_X(260) = 0.35$  from which the required probabilities and percentages are obtained straightforwardly.

### Important Mathematical Characteristics

In addition to inheriting all the mathematical characteristics of the Gaussian random variable, the standard normal random variable in its own right has the following relationship to the Chi-square random variable: If  $X \sim N(0, 1)$  then  $X^2 \sim \chi^2(1)$  (a result that was actually established in Chapter 6, Example 6.3). In general, if  $X \sim N(\mu, \sigma^2)$ , then  $[(X - \mu)/\sigma]^2 \sim \chi^2(1)$ . Finally, by virtue of the reproductive property of the Chi-square random variable, if  $X_i; i = 1, 2, \dots, n$  are  $n$  independent Gaussian random variables with identical distributions  $N(\mu, \sigma^2)$ , then the random variable defined as:

$$Y = \sum_{i=1}^n \left( \frac{X_i - \mu}{\sigma} \right)^2 \quad (9.138)$$

possesses a  $\chi^2(n)$  distribution (See Exercise 9.20). These results find extensive application in statistical inference.

#### 9.2.3 The Lognormal Random Variable

##### Basic Characteristics and Model Development

By analogy with the Gaussian random variable, consider a random variable whose observed value is composed of a *product* of many small independent

random quantities, i.e.

$$X = \prod_{i=1}^n X_i \quad (9.139)$$

Taking natural logarithms of this expression yields:

$$\ln X = \sum_{i=1}^n \ln X_i = \sum_{i=1}^n Y_i \quad (9.140)$$

Following the discussion in the previous sections, we now know that in the limit as  $n$  becomes very large,  $Y = \sum_{i=1}^n Y_i$  tends to behave like a Gaussian random variable, with the implication that  $\ln X$  is a random variable with a Gaussian (Normal) pdf. If the mean is designated as  $\alpha$  and variance as  $\beta^2$ , then the pdf for the random variable  $Y = \ln X$  is:

$$g(y) = \frac{1}{\beta\sqrt{2\pi}} \exp\left\{\frac{-(y-\alpha)^2}{2\beta^2}\right\} \quad (9.141)$$

Using techniques discussed in Chapter 6, it can be shown (see Exercise 9.21) that from the variable transformation and its inverse,

$$\begin{aligned} Y &= \ln X \\ X &= e^Y \end{aligned} \quad (9.142)$$

one obtains from Eq (9.141) the required pdf for  $X$  as:

$$f(x) = \frac{1}{x\beta\sqrt{2\pi}} \exp\left\{\frac{-(\ln x - \alpha)^2}{2\beta^2}\right\}; 0 < x < \infty \quad (9.143)$$

The random variable  $X$  whose pdf is shown above, is referred to as a lognormal random variable for the obvious reason that the (natural) logarithm of  $X$  has a “normal” distribution. Eqn (9.143) is therefore an expression of the pdf for the lognormal random variable with parameters  $\alpha$  and  $\beta$ , i.e.  $X \sim \mathcal{L}(\alpha, \beta^2)$ .

An alternative form of this pdf is obtained by defining

$$\alpha = \ln m; \Rightarrow m = e^\alpha \quad (9.144)$$

so that Eq (9.143) becomes

$$f(x) = \frac{1}{x\beta\sqrt{2\pi}} \exp\left\{\frac{-(\ln x/m)^2}{2\beta^2}\right\}; 0 < x < \infty \quad (9.145)$$

### Important Mathematical Characteristics

The following are some key mathematical characteristics of the  $\mathcal{L}(\alpha, \beta^2)$  random variable and its pdf:

1. **Characteristic parameters:**  $\alpha$ , (or  $m > 0$ ) is the *scale* parameter;  $\beta > 0$ , (or  $w = e^{\beta^2}$ ) is the *shape* parameter. Because of the structural similarity between this pdf and the Gaussian pdf, it is easy to misconstrue  $\alpha$  as the mean of the random variable  $X$ , and  $\beta$  as the standard deviation. The reader must be careful not to fall into this easy trap:  $\alpha$  is the expected value not of  $X$ , but of  $\ln X$ ; and  $\beta^2$  is the variance of  $\ln X$ ; i.e.

$$\begin{aligned} E(\ln X) &= \alpha \\ \text{Var}(\ln X) &= \beta^2 \end{aligned} \quad (9.146)$$

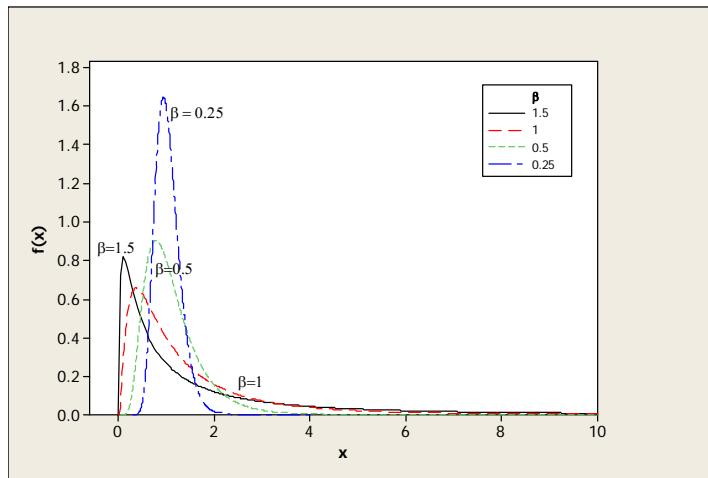
It is precisely for this reason that in this textbook we have deliberately opted not to use  $\mu$  and  $\sigma$  in the lognormal pdf: using these symbols leads to too much confusion.

2. **Mean:**  $E(X) = \exp(\alpha + \beta^2/2) = m e^{\beta^2/2} = m \sqrt{w}$ ;  
 $\text{Mode} = m/w = e^{(\alpha-\beta^2)}$   
 $\text{Median} = m = e^\alpha$
3. **Variance:**  $\text{Var}(X) = e^{(2\alpha+\beta^2)} (e^{\beta^2} - 1) = m^2 w (w - 1)$   
Note that  $\text{Var}(X) = [E(X)]^2 (w - 1)$
4. **Higher Moments:** Coefficient of Skewness:  $\gamma_3 = (w + 2)\sqrt{(w - 1)}$ ;  
Coefficient of Kurtosis:  $\gamma_4 = w^4 + 2w^3 + 3w^2 - 3$ .
5. **Moment generating and Characteristic functions:** Even though all moments exist for the lognormal distribution, the MGF does not exist. The characteristic function exists but is quite complicated.

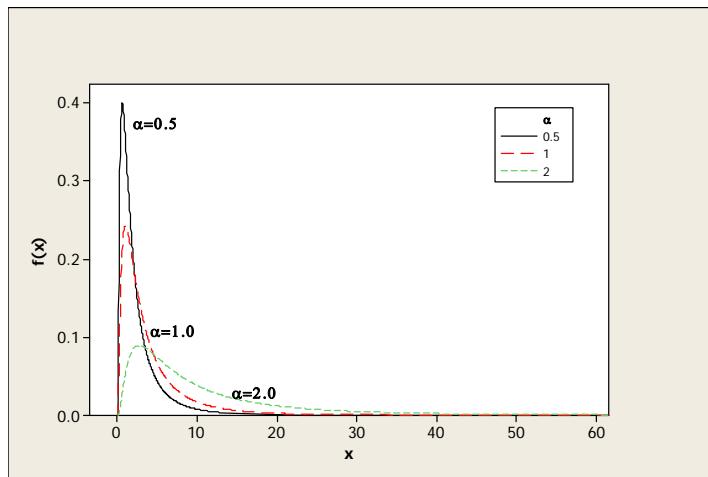
A plot of the lognormal distribution for various values of the shape parameter  $\beta$ , with the scale parameter fixed at  $\alpha = 0$ , is shown in Fig 9.8. On the other hand, a plot for various values of  $\alpha$ , with the shape parameter fixed at  $\beta = 1$ , is shown in Fig 9.9.

An important point that must not be missed here is as follows: whereas for the Gaussian distribution  $\mu$  is a location parameter responsible for “shifting” the distribution, the corresponding parameter for the lognormal distribution,  $\alpha$ , does *not* shift the distribution’s location but rather *scales* its magnitude. This is consistent with the fact that the “additive” characteristics underlying the Gaussian distribution correspond to “multiplicative” characteristics in the lognormal distribution. Thus, while a change in  $\mu$  shifts the location of the Gaussian distribution, a change in  $\alpha$  scales (by multiplication) the lognormal distribution. The parameter  $\alpha$  is a location parameter *only* for the distribution of  $\ln X$  (which is Gaussian) *not* for the distribution of  $X$ .

A final point to note: while the most popular measure of central location, the mean  $E(X)$ , depends on *both*  $\alpha$  and  $\beta$  for the lognormal distribution, the median on the other hand,  $m = e^\alpha$ , depends only on  $\alpha$ . This suggests that the median is a more natural indicator of central location for the lognormal



**FIGURE 9.8:** Lognormal pdfs for scale parameter  $\alpha = 0$  and various values of the shape parameter  $\beta$ . Note how the shape changes, becoming less skewed as  $\beta$  becomes smaller.



**FIGURE 9.9:** Lognormal pdfs for shape parameter  $\beta = 1$  and various values of the scale parameter  $\alpha$ . Note how the shape remains unchanged while the entire distribution is scaled appropriately depending on the value of  $\alpha$ .

random variable. By the same token, a more natural measure of dispersion for this random variable is  $Var(X)/[E(x)]^2$ , the variance scaled by a square of the mean value, or, equivalently, the square of  $C_v$ , the coefficient of variation: from the expression given above for the variance, this quantity is  $(w - 1)$ , depending only on  $w = e^{\beta^2}$ .

### Applications

From the preceding considerations regarding the genesis of the lognormal distribution, it is not surprising that the following phenomena generate random variables that are well-modeled with the lognormal pdf:

1. Size of particles obtained by breakage (grinding) or granulation of fine particles;
2. Size of living organisms, especially where growth depends on numerous factors proportional to instantaneous organism size;
3. Personal income, or net worth, or other such quantities for which the current observation is a random proportion of the previous value (e.g., closing price of stocks, or index options).

The lognormal distribution is therefore used for describing particle size distributions in mining and granulation processes as well as in atmospheric studies; for molecular weight distributions of polymers arising from complex reaction kinetics; for distributions of incomes in a free market economy.

Because it is a skewed distribution just like the gamma distribution, the lognormal distribution is sometimes used to describe such phenomena as latent periods of infectious diseases, or age at the onset of such diseases as Alzheimer's or arthritis — phenomena that are more naturally described by the gamma density since they involve the time to the occurrence of events driven by multiple cumulative effectors.

Finally, in statistical inference applications, probabilities are traditionally computed for lognormal random variables using Normal probability tables. Observe that if  $X \sim \mathcal{L}(\alpha, \beta^2)$ , then:

$$P(a_1 < X < a_2) = P(\ln a_1 < Y < \ln a_2) \quad (9.147)$$

where  $Y \sim N(\alpha, \beta^2)$ . Thus, using tables of standard normal cumulative probabilities, one is able to obtain from Eq (9.147) that:

$$\begin{aligned} P(a_1 < X < a_2) &= P\left[\left(\frac{\ln a_1 - \alpha}{\beta}\right) < Z < \left(\frac{\ln a_2 - \alpha}{\beta}\right)\right] \\ &= F\left(\frac{\ln a_2 - \alpha}{\beta}\right) - F\left(\frac{\ln a_1 - \alpha}{\beta}\right) \end{aligned} \quad (9.148)$$

However, statistical software packages such as MINITAB provide more efficient means of computing desired probabilities directly without resorting to tables based on such transformations.

**Example 9.4 PRODUCT QUALITY ATTAINMENT IN GRANULATION PROCESS**

Granular products made from pan granulation processes are typically characterized by their particle size distributions (and bulk densities). Material produced in a particular month at a manufacturing site has a particle size distribution that is well-characterized by a lognormal distribution with scale parameter  $\alpha = 6.8$  and shape parameter  $\beta = 0.5$ , i.e.  $X \sim \mathcal{L}(6.8, 0.5)$ . The product quality requirements (in microns) is specified as  $350 \mu\text{m} < X < 1650 \mu\text{m}$ , and the *yield* of the process is the percentage of the product meeting this requirement. To be profitable, the manufacturing process yield must be *at least* 85% month-to-month.

(i) What is this month's process yield? (ii) If particles for which  $X < 350 \mu\text{m}$  are classified as "fines" and can be recycled, how much of the material made during this month falls into this category?

**Solution:**

The problem requires computing the following probabilities: (i)  $P(350 < X < 1650)$  and (ii)  $P(X < 350)$  for the  $\mathcal{L}(6.8, 0.5)$  distribution. These values are obtained directly from MINITAB (or other such software) directly without using transformations and tables of standard normal cumulative probabilities. First,

$$P(350 < X < 1650) = 0.858 \quad (9.149)$$

as shown in Fig (9.10). This indicates a yield of 85.8%, the percentage of material produced meeting the quality requirement; the manufacturing plant is therefore profitable this particular month.

(ii) The second required probability is obtained directly from MINITAB as  $P(X < 350) = 0.030$ . Thus, 3% of the material will be classified as "fines" that can be recycled.

These two probabilities imply that there is a residual 11.2% of the product with size in excess of 1650 microns, falling into the "oversized" category, since by definition:

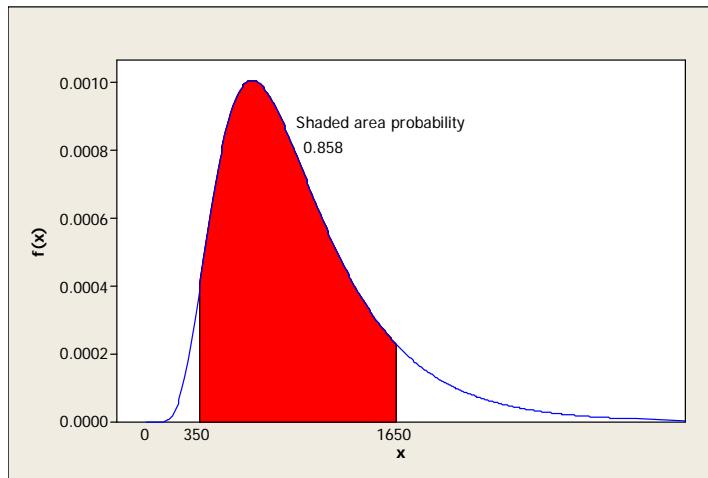
$$1 = P(X < 350) + P(350 < X < 1650) + P(X > 1650) \quad (9.150)$$

#### 9.2.4 The Rayleigh Random Variable

Let us consider a 2-dimensional vector  $(X_1, X_2)$  where the components are mutually independent random variables representing random deviations of hits from a target on a plane whose coordinates are  $X_1, X_2$ . The magnitude of this random vector is the random variable  $X$ ,

$$X = \sqrt{X_1^2 + X_2^2} \quad (9.151)$$

In such a case,  $X$  is known as a Rayleigh random variable. To obtain the probability model for this random variable, we merely need to recall that this



**FIGURE 9.10:** Particle size distribution for the granulation process product: a lognormal distribution with  $\alpha = 6.8, \beta = 0.5$ . The shaded area corresponds to product meeting quality specifications,  $350 < X < 1650$  microns.

description is exactly as in the Herschel/Maxwell model presented earlier, except that this time

$$x = \sqrt{r^2} \quad (9.152)$$

Thus, from Eq (9.120), we obtain immediately that,

$$f(x) = \frac{x}{b^2} e^{-\frac{1}{2} \frac{x^2}{b^2}}; x > 0; b > 0 \quad (9.153)$$

This is the pdf for the Rayleigh random variable  $\mathcal{R}(b)$ . It can be shown via methods presented in Chapter 6 that if  $Y_1 \sim N(0, b^2)$  and  $Y_2 \sim N(0, b^2)$  then

$$X = \sqrt{Y_1^2 + Y_2^2} \quad (9.154)$$

possesses a Rayleigh  $\mathcal{R}(b)$  distribution (See Exercise 9.25).

### Important Mathematical Characteristics

The following are some key mathematical characteristics of the  $\mathcal{R}(b^2)$  random variable and its pdf:

1. **Characteristic parameter:**  $b > 0$  is the *scale* parameter;
2. **Mean:**  $E(X) = b\sqrt{\pi/2}$ ; Mode =  $b$ ; Median =  $b\sqrt{\ln 4}$
3. **Variance:**  $Var(X) = b^2(2 - \pi/2)$

**4. Higher Moments:**

Coefficient of Skewness:  $\gamma_3 = 2\sqrt{\pi}(\pi - 3)/(4 - \pi)^{3/2}$ ;  
 Coefficient of Kurtosis:  $\gamma_4 = 3 - [(6\pi^2 - 24\pi + 16)/(\pi^2 - 8\pi + 16)]$ .

**5. Moment generating and Characteristic functions:** Both the MGF and CF exist but are quite complicated.

As a final point of interest, we note that the Weibull pdf with  $\zeta = 2$  is identical to a Rayleigh pdf with the parameter  $b = \sqrt{\beta/2}$ . At first blush, this appears to be no more than an odd coincidence since, from the preceding discussions, there is no obvious physical reason for the Weibull distribution, which is most appropriate for reliability and life-testing problems, to encompass as a special case, the Rayleigh distribution. In other words, there is no apparent structural reason for a member of the Gaussian family (the Rayleigh random variable) to be a special case of the Weibull distribution, a member of the totally different Gamma family. However there are two rational justifications for this surprising connection: the first purely “empirical”, the second more structural.

1. Recall Eq (9.52) where we noted that the Weibull pdf arose from a specification of a generic cumulative hazard function (chf) made deliberately more complex than that of the exponential by simply introducing a power: any other conceivable (and differentiable) chf,  $H(t)$ , will give rise to a pdf given by Eq (9.52). In this case, the Rayleigh random variable arises by specifying a specific chf (in terms of  $x$ , rather than time,  $t$ ),  $H(x) = (\eta x)^2$ . However, this perspective merely connects the Rayleigh to the Weibull distribution through the chf; it gives no insight into why the specific choice of  $\zeta = 2$  is a special case of such interest as to represent an entirely unique random variable.
2. For structural insight, we return to the exponential pdf and view it (as we are perfectly at liberty to do) as a distribution of distances of “particles” from a fixed point in 1-dimension (for example, the length to the discovery of the first flaw on a length of fiber-optic cable), where the “particles” (e.g. “flaws”) occur with uniform Poisson intensity  $\eta$ . From this perspective, the hazard function  $h(x) = \eta$  (with the corresponding linear  $H(t) = \eta t$ ) represents the uniform exponential random variable “failure rate,” if failure is understood as “not finding a flaw” at a location between  $x$  and  $x + \Delta x$ . The 2-dimensional version of this problem, the distribution of the radial distances of “flaws” from a fixed center point in a plane (where  $X^2 = (X_1^2 + X_2^2)$ ) has a square form for the chf,

$$H(x) = (\eta x)^2 \quad (9.155)$$

with a corresponding hazard function  $h(x) = 2\eta(\eta x)$  indicating that the “failure rate” (the rate of “not finding a flaw” at a radial distance  $x$  from the center in a 2-dimensional plane) increases with  $x$ . This, of course, is

precisely the conceptual model used to derive the Rayleigh pdf; it shows how  $\zeta = 2$  in the Weibull distribution is structurally compatible with the phenomenon underlying the Rayleigh random variable.

## Applications

The Rayleigh distribution finds application in military studies of battle damage assessment, especially in analyzing the distance of bomb hits from desired targets (not surprising given the discussion above). The distribution is also used in communication theory for modeling communication channels, and for characterizing satellite Synthetic Aperture Radar (SAR) data.

### 9.2.5 The Generalized Gaussian Model

To conclude, we note that all three random variables discussed above can be represented as special cases of the random variable  $X$  with the following pdf:

$$f(x) = C_1(x) \exp \left\{ \frac{-(C_2(x) - C_3)^2}{2C_4} \right\} \quad (9.156)$$

- 1. Gaussian (Normal):  $C_1(x) = 1/(\sigma\sqrt{2\pi})$ ;  $C_2(x) = x$ ;  $C_3 = \mu$ ;  $C_4 = \sigma^2$ ;
- 2. Lognormal:  $C_1(x) = 1/(x\beta\sqrt{2\pi})$ ;  $C_2(x) = \ln x$ ;  $C_3 = \alpha$ ;  $C_4 = \beta^2$ ;
- 3. Rayleigh:  $C_1(x) = x/b^2$ ;  $C_2(x) = x$ ;  $C_3 = 0$ ;  $C_4 = b^2$ ;

---

## 9.3 Ratio Family Random Variables

The final family of continuous random variables to be discussed in this Chapter consists of the following 5 members:

- The Beta random variable,
- The (Continuous) Uniform random variable,
- Fisher's  $F$  random variable,
- Student's  $t$  random variable, and
- The Cauchy random variable

This grouping of random variables is far more diverse than any of the earlier two groupings. From the first two that are defined only on bounded regions of finite size on the real line, to the last two that are always symmetric and are defined on the entire real line, and the third one that is defined only on the semi-infinite positive real line, these random variables appear to have nothing in common. Nevertheless, all members of this group do in fact share a very important common characteristic: as the "family name" implies, they all arise as ratios composed of other (previously encountered) random variables.

Some of the most important random variables in statistical inference belong to this family; and one of the benefits of the upcoming discussion is that the ratios from which they arise provide immediate indications (and justification) of the role each random variable plays in statistical inference. In the interest of limiting the length of a discussion that is already quite long, we will simply state the results, suppressing the derivation details entirely, or else referring the reader to appropriate places in Chapter 6 where we had earlier provided such derivations in anticipation of these current discussions.

### 9.3.1 The Beta Random Variable

#### Basic Characteristics and Model Development

Consider two mutually independent gamma random variables  $Y_1$  and  $Y_2$  possessing  $\gamma(\nu, 1)$  and  $\gamma(\omega, 1)$  pdf's respectively. Now define a new random variable  $X$  as follows:

$$X = \frac{Y_1}{Y_1 + Y_2} \quad (9.157)$$

i.e. the proportion contributed by  $Y_1$  to the ensemble sum. Note that  $0 < x < 1$ , so that the values taken by this random variable will be fractional. A random variable thus defined is a beta random variable, and we are now interested in obtaining its pdf.

From the pdfs of the gamma random variables, i.e.

$$f(y_1) = \frac{1}{\Gamma(\nu)} y_1^{\nu-1} e^{-y_1}; 0 < y_1 < \infty \quad (9.158)$$

$$f(y_2) = \frac{1}{\Gamma(\omega)} y_2^{\omega-1} e^{-y_2}; 0 < y_2 < \infty \quad (9.159)$$

we may use methods discussed in Chapter 6 (specifically, see Example 6.3) to obtain that the required pdf for  $X$  defined in Eq (9.157), is:

$$f(x) = \frac{\Gamma(\nu + \omega)}{\Gamma(\nu)\Gamma(\omega)} x^{\nu-1} (1-x)^{\omega-1}; 0 < x < 1; \nu > 0; \omega > 0 \quad (9.160)$$

### The Model and Some Remarks

The pdf for  $X$ , the Beta random variable,  $B(\alpha, \beta)$ , is:

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}; 0 < x < 1; \alpha > 0; \beta > 0 \quad (9.161)$$

The name arises from the relationship between the pdf above and the Beta function defined by:

$$\text{Beta}(\alpha, \beta) = \int_0^1 u^{\alpha-1} (1-u)^{\beta-1} du = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \quad (9.162)$$

This pdf in Eq (9.161) is the first continuous model to be restricted to a finite interval, in this case,  $x \in [0, 1]$ . This pdf is defined on this unit interval, but as we show later, it is possible to generalize it to a pdf on an arbitrary finite interval  $[\delta_0, \delta_1]$ .

### Important Mathematical Characteristics

The following are some key mathematical characteristics of the  $B(\alpha, \beta)$  random variable and its pdf:

1. **Characteristic parameters:**  $\alpha > 0$  and  $\beta > 0$  are *both* shape parameters. The pdf takes on a wide variety of shapes depending on the values of these two parameters. (See below.)
2. **Mean:**  $E(X) = \alpha/(\alpha + \beta)$ ;  
**Mode** =  $(\alpha - 1)/(\alpha + \beta - 2)$  for  $\alpha > 1, \beta > 1$ , otherwise no mode exists.  
There is no closed form expression for the Median.
3. **Variance:**  $Var(X) = \alpha\beta/(\alpha + \beta)^2(\alpha + \beta + 1)$
4. **Higher Moments:**  
Coefficient of Skewness:

$$\gamma_3 = \frac{2(\beta - \alpha)}{(\alpha + \beta + 2)} \sqrt{\left( \frac{1}{\alpha} + \frac{1}{\beta} + \frac{1}{\alpha\beta} \right)}$$

Coefficient of Kurtosis:

$$\gamma_4 = \frac{3(\alpha + \beta)(\alpha + \beta + 1)(\alpha + 1)(2\beta - \alpha)}{\alpha\beta(\alpha + \beta + 2)(\alpha + \beta + 3)} + \frac{\alpha(\alpha - \beta)}{(\alpha + \beta)}$$

5. **Moment generating and Characteristic functions:** Both the MGF and CF exist but are quite complicated.

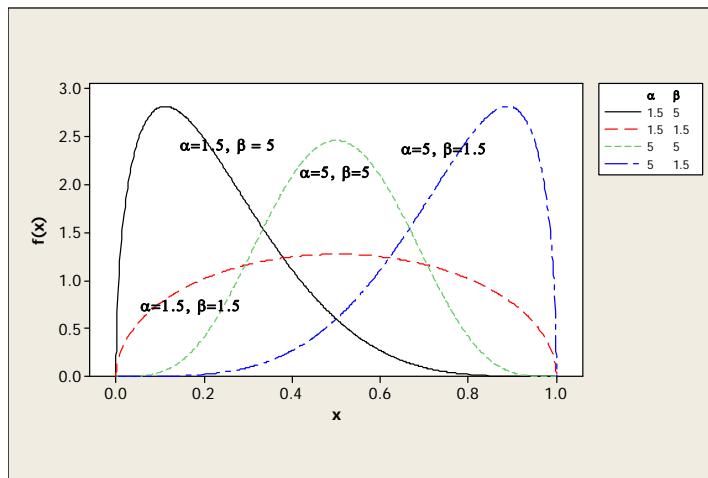
### The Many Shapes of the Beta Distribution

First, here are a few characteristics that can be deduced directly from an examination of the functional form of  $f(x)$ :

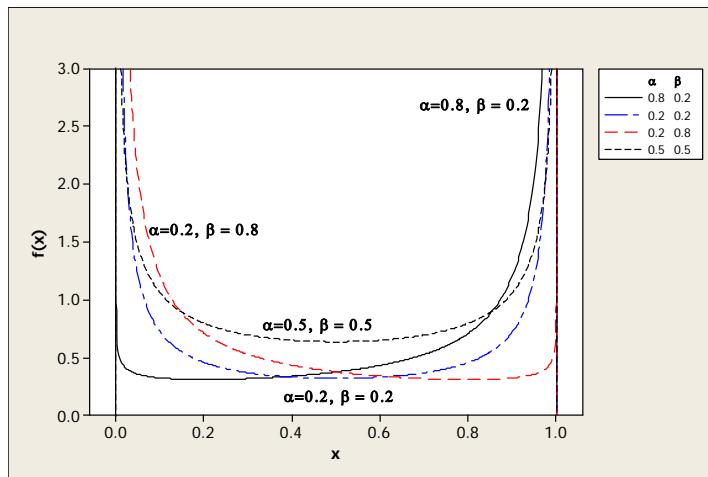
1. For  $\alpha > 1$  and  $\beta > 1$ , the powers to which  $x$  and  $(1 - x)$  are raised will be positive; as a result,  $f(x) = 0$  at both boundaries  $x = 0$  and  $x = 1$  and will therefore have to pass through a maximum somewhere in the interval  $(0,1)$ , at a location determined by the value of  $\alpha$  relative to  $\beta$ .
2. Conversely, for  $0 < \alpha, \beta < 1$  the powers to which  $x$  and  $(1 - x)$  are raised will be negative so that  $f(x)$  has asymptotes at both boundaries  $x = 0$  and  $x = 1$ .
3. For  $0 < \alpha < 1$  and  $\beta > 1$ ,  $f(x)$  has an asymptote at  $x = 0$  but is zero at  $x = 1$ ; complementarily, for  $0 < \beta < 1$  and  $\alpha > 1$ ,  $f(x)$  has an asymptote at  $x = 1$  and is zero at  $x = 0$ .
4. For  $\alpha = \beta$ ,  $f(x)$  is symmetric, with  $\alpha = \beta = 1$  being a special case in which  $f(x)$  is flat — a case of special significance warranting a separate discussion.

From these considerations, the following observations of the shapes of the Beta distribution follow:

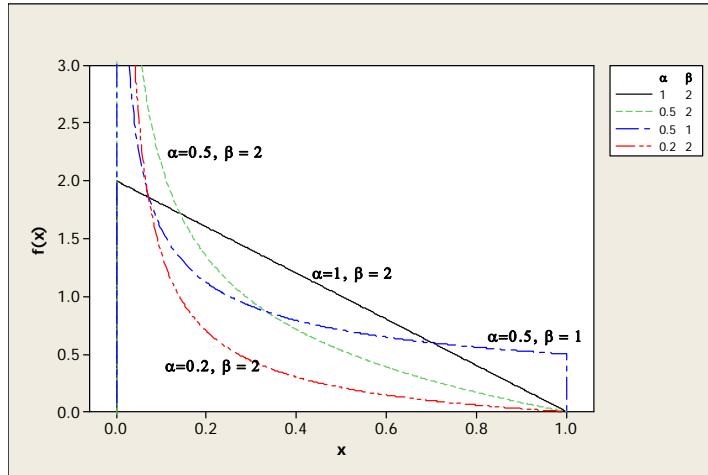
1. When  $\alpha, \beta > 1$ , the Beta distribution is unimodal, skewed left when  $\alpha > \beta$ , skewed right when  $\alpha < \beta$ , and symmetric when  $\alpha = \beta$ , as shown in Fig 9.11.
2. When  $\alpha, \beta < 1$ , the Beta distribution is “U-shaped” with a sharper approach to the *left* asymptote at  $x = 0$  when  $\alpha > \beta$ , a sharper approach to the *right* asymptote at  $x = 1$  when  $\alpha < \beta$ , and a symmetric U-shape when  $\alpha = \beta$ , as shown in Fig 9.12.
3. When  $(\alpha - 1)(\beta - 1) \leq 0$ , the Beta distribution is “J-shaped” with a left-handed “J” when  $\alpha < \beta$ , ending at zero at  $x = 1$  for  $\beta \neq 1$  and at a non-zero value when  $\beta = 1$  (as shown in Fig 9.13). The Beta distribution is a right-handed “J” when  $\alpha > \beta$  (not shown).
4. In the special case when  $\alpha = 1$  and  $\beta = 2$ , the Beta distribution is a right-sloping straight line, as shown in Fig 9.13; when  $\alpha = 2$  and  $\beta = 1$  we obtain a left-sloping straight line (not shown).



**FIGURE 9.11:** Unimodal Beta pdfs when  $\alpha > 1; \beta > 1$ : Note the symmetric shape when  $\alpha = \beta$ , and the skewness determined by the value of  $\alpha$  relative to  $\beta$



**FIGURE 9.12:** U-Shaped Beta pdfs when  $\alpha < 1; \beta < 1$



**FIGURE 9.13:** Other shapes of the Beta pdfs: It is J-shaped when  $(\alpha - 1)(\beta - 1) < 0$  and a straight line when  $\beta = 2; \alpha = 1$

### Applications

The Beta distribution naturally provides a good model for many random phenomena involving proportions. For example, it is used in Bayesian analysis (see later) for describing *a-priori* knowledge about  $p$ , the Binomial probability of success. Another example practical application (mostly in quality assurance) arises from the following result:

Given  $n$  independent random observations,  $\omega_1, \omega_2, \dots, \omega_n$  from a phenomenon possessing an arbitrary pdf, rank the observations from the smallest to the largest as  $y_1, y_2, \dots, y_n$  (i.e. with  $y_1$  as the smallest and  $y_n$  as the largest). If  $y_r$  is the  $r^{th}$ -smallest and  $y_{n-s+1}$  is the  $s^{th}$ -largest, then regardless of the underlying pdf of the variable,  $X$ , the proportion of the population between  $y_r$  and  $y_{n-s+1}$  possesses a  $B(\alpha, \beta)$  distribution with  $\alpha = (n - s + 1) - r$ , and  $\beta = r + s$ , i.e.

$$f(x; n, r, s) = \frac{\Gamma(n+1)}{\Gamma(n-r-s+1)\Gamma(r+s)} x^{n-r-s} (1-x)^{r+s-1}; \quad (9.163)$$

This important result frees one from making any assumptions about the underlying pdf of the population from which the original quality data  $\omega_1, \omega_2, \dots, \omega_n$  came from.

The example below illustrates yet another application of the Beta distribution, in functional genomics studies.

#### Example 9.5 DIFFERENTIAL GENE EXPRESSION STATUS FROM MICROARRAYS

In functional genomics, one of the objectives is to provide a quantitative (as opposed to qualitative) understanding of the functions of genes

and how they regulate the function of complex living organisms. The advent of the high-throughput microarray technology has made it possible to collect expression data on every gene in a cell simultaneously. Such microarray data, usually presented in the form of fluorescence signal intensities measured from each spot  $i$  on a microarray, yield an ordered pair  $(y_{i_1}, y_{i_0})$  with  $y_{i_1}$  coming from the gene in question under *test* conditions (e.g. from a cancer cell), and  $y_{i_0}$  under control conditions (e.g. normal cell). In theory, if the gene is up-regulated under test conditions,  $FC = y_{i_1}/y_{i_0} > 1$ ,  $FC < 1$  if down-regulated, and  $FC = 1$  if non-differentially regulated. This quantity,  $FC$ , is the so-called “fold-change” associated with this gene under the test conditions. However, because of measurement noise and other myriad technical considerations, this ratio is difficult to characterize statistically and not terribly reliable by itself.

It has been suggested in Gelmi<sup>6</sup>, to use the fractional intensity  $x_i$  defined by:

$$x_i = \frac{y_{i_1}}{y_{i_1} + y_{i_0}} \quad (9.164)$$

because this fraction can be shown to possess a Beta distribution. In this case, theoretically,  $x_i > 0.5$  for up-regulated genes,  $x_i < 0.5$  for down-regulated genes, and  $x_i = 0.5$  for non-differentially regulated ones, with noise introducing uncertainties into the computed values. Microarray data in the form of  $x_i$  may thus be analyzed using the Beta distribution to provide probabilities that the gene in question is up-, down-, or non-differentially-regulated.

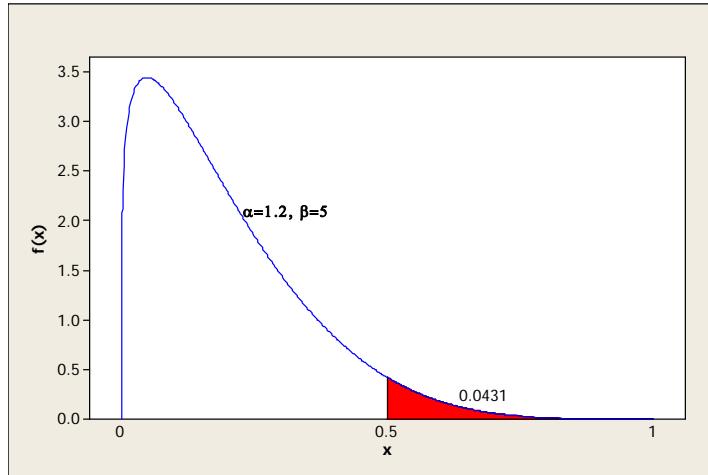
Replicate data on a particular gene of interest on a microarray yielded an average fractional intensity value  $\bar{x} = 0.185$  and standard deviation  $s = 0.147$ . Using estimation techniques discussed in Part IV, it has been determined that this result indicates that the fractional intensity data can be considered as a sample from a theoretical Beta distribution with  $\alpha = 1.2$ ;  $\beta = 5.0$ . Determine the probability that the gene is up-regulated.

#### **Solution:**

Recalling that if a gene in question is truly up-regulated,  $x_i > 0.5$ , the problem requires computing  $P(X_i > 0.5)$ . This is obtained directly from MINITAB as 0.0431 (See Fig 9.14), indicating that it is highly unlikely that this particular gene is in fact up-regulated.

---

<sup>6</sup>Gelmi, C. A. (2006) “A novel probabilistic framework for microarray data analysis: From fundamental probability models to experimental validation”, PhD Thesis, University of Delaware.



**FIGURE 9.14:** Theoretical distribution for characterizing fractional microarray intensities for the example gene: The shaded area corresponds to the probability that the gene in question is upregulated.

### 9.3.2 Extensions and Special Cases of the Beta Random Variable

#### Generalized Beta Random Variable

It is relatively easy to generalize the Beta pdf to cover the interval  $(\delta_0, \delta_1)$  where  $\delta_0$  and  $\delta_1$  are not necessarily 0 and 1, respectively. The general pdf is:

$$f(x) = \frac{1}{(\delta_1 - \delta_0)} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \left( \frac{x - \delta_0}{\delta_1 - \delta_0} \right)^{\alpha-1} \left( 1 - \frac{x - \delta_0}{\delta_1 - \delta_0} \right)^{\beta-1} \quad (9.165)$$

$\delta_0 < x < \delta_1; \alpha > 0; \beta > 0$

#### Inverted Beta Random Variable

Let the random variable  $X_1$  have a  $B(\alpha, \beta)$  pdf and define a new random variable as

$$X = \frac{X_1}{1 - X_1} \quad (9.166)$$

It is easy to show from the preceding discussion (on the genesis of the Beta random variable from contributing gamma random variables  $Y_1$  and  $Y_2$ ) that  $X$  defined as in Eq (9.166) above may also be represented as:

$$X = \frac{Y_1}{Y_2} \quad (9.167)$$

i.e. as a ratio of gamma random variables. The random variable  $X$  defined in this manner has the Inverted Beta pdf:

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{x^{\alpha-1}}{(1+x)^{\alpha+\beta}}; x > 0; \alpha > 0; \beta > 0 \quad (9.168)$$

This is sometimes also referred to by the somewhat more cumbersome “Beta distribution of the second kind”. This random variable is related to the  $F$ -distribution to be discussed shortly.

### 9.3.3 The (Continuous) Uniform Random Variable

#### Basic Characteristics, Model and Remarks

When  $\alpha = \beta = 1$  in the Beta pdf, the result is the special function:

$$f(x) = 1; 0 < x < 1 \quad (9.169)$$

a pdf of a random variable that is uniform on the interval  $(0,1)$ . The general uniform random variable,  $X$ , defined on the interval  $(a, b)$  has the pdf:

$$f(x) = \frac{1}{b-a}; a < x < b \quad (9.170)$$

It is called a uniform random variable,  $U(a, b)$ , for the obvious reason that, unlike all the other distributions discussed thus far, the probability description for this random variable is completely uniform, favoring no particular value in the specified range. The uniform random variable on the unit interval  $(0,1)$  is known as a standard uniform random variable.

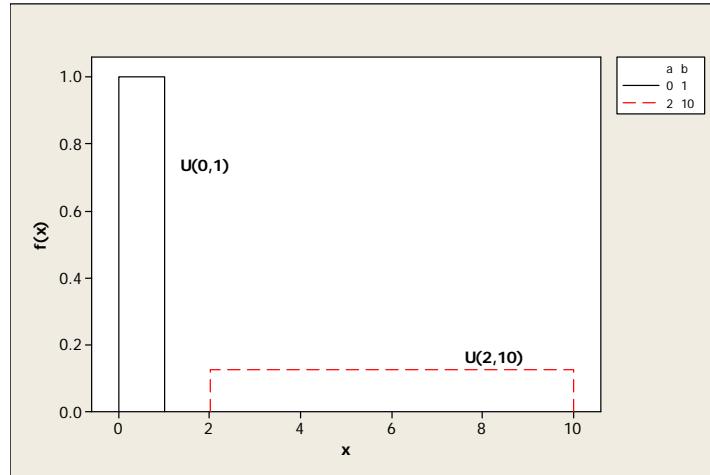
#### Important Mathematical Characteristics

The following are some key mathematical characteristics of the  $U(a, b)$  random variable and its pdf:

1. **Characteristic parameters:**  $a, b$  jointly form the range, with  $a$  as the location parameter (see Fig 9.15). Narrower distributions are longer than wider ones because the total probability area must equal 1 in every case.
2. **Mean:**  $E(X) = (a + b)/2$ ;  
Median = Mean;  
Mode: non-unique; all values in interval  $(a, b)$ ;
3. **Variance:**  $Var(X) = (b - a)^2/12$
4. **Moment generating and Characteristic functions:**

$$M(t) = \frac{e^{bt} - e^{at}}{t(b-a)};$$

$$\varphi(t) = \frac{e^{jbt} - e^{jat}}{jt(b-a)}$$



**FIGURE 9.15:** Two uniform distributions over different ranges  $(0,1)$  and  $(2,10)$ . Since the total area under the pdf must be 1, the narrower pdf is proportionately longer than the wider one.

Of the direct relationships between the  $U(a, b)$  random variable and other random variables, the most important are (i) If  $X$  is a  $U(0, 1)$  random variable, then  $Y = -\beta \ln X$  is an exponential random variable  $\mathcal{E}(\beta)$ . This result was established in Example 6.2 in Chapter 6. (ii) If  $X \sim U(0, 1)$  then  $Y = 1 - X^{1/\beta}$  is a Beta random variable  $B(1, \beta)$ .

### Applications

The uniform pdf is the obvious choice for describing equiprobable events on bounded regions of the real-line. (The discrete version is used for equiprobable discrete events in a sample space.) But perhaps its most significant application is for generating random numbers for other distributions.

#### 9.3.4 Fisher's F Random Variable

##### Basic Characteristics, Model and Remarks

Consider two mutually independent random variables  $Y_1$  and  $Y_2$  respectively possessing  $\chi^2(\nu_1)$  and  $\chi^2(\nu_2)$  distributions; a random variable  $X$  defined as:

$$X = \frac{Y_1/\nu_1}{Y_2/\nu_2} \quad (9.171)$$

is a Fisher  $F$  random variable,  $F(\nu_1, \nu_2)$  named in honor of R. A. Fisher (1890-1962), the father of the field of Statistical Design of Experiments. It is possible

to show that its pdf is:

$$f(x) = \frac{\Gamma[\frac{1}{2}(\nu_1 + \nu_2)]}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \frac{\nu_1^{\nu_1/2}\nu_2^{\nu_2/2}}{(\nu_1 x + \nu_2)^{(\nu_1+\nu_2)/2}}; 0 < x < \infty \quad (9.172)$$

### Important Mathematical Characteristics

The following are some key mathematical characteristics of the  $F(\nu_1, \nu_2)$  random variable and its pdf:

1. **Characteristic parameters:**  $\nu_1, \nu_2$ , parameters inherited directly from the contributing Chi-square random variables, retain their Chi-square distribution characteristics as “degrees of freedom.”
2. **Mean:**

$$E(X) = \mu = \frac{\nu_2}{\nu_2 - 2}; \nu_2 > 2$$

#### Mode:

$$x^* = \frac{\nu_2(\nu_1 - 2)}{\nu_1(\nu_1 + 2)}; \nu_1 > 2$$

3. **Variance:**

$$Var(X) = \sigma^2 = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}; \nu_2 > 4 \quad (9.173)$$

Expressions for Skewness and Kurtosis are a bit complicated; the MGF does not exist and the expression for the CF is complicated.

Figure 9.16 shows two  $F$  distributions for the same value of  $\nu_2 = 15$  but different values of  $\nu_1$ .

The  $F$  distribution is related to two additional pdf's as follows: If  $X$  has an  $F$  distribution with  $\nu_1, \nu_2$  degrees of freedom, then

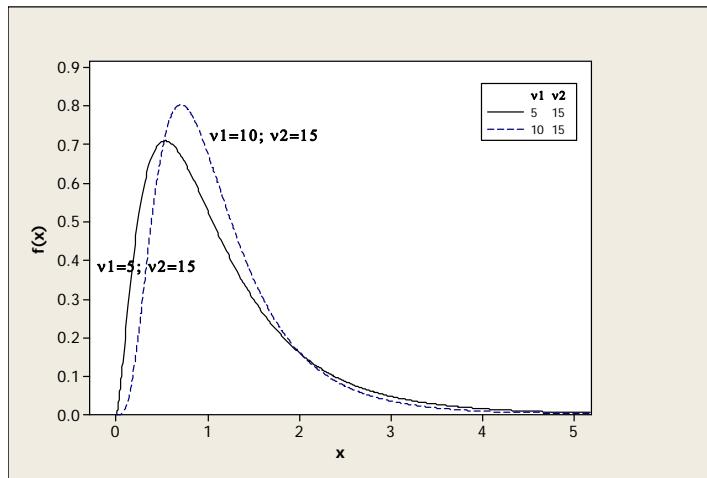
$$Y = \frac{(\nu_1/\nu_2) X}{1 + (\nu_1/\nu_2) X} \quad (9.174)$$

has a Beta  $B(\nu_1/2, \nu_2/2)$  distribution.

Alternatively, if  $Y$  has an Inverted Beta distribution with  $\alpha = \nu_1/2; \beta = \nu_2/2$  then

$$X = \frac{\nu_1}{\nu_2} Y \quad (9.175)$$

has an  $F(\nu_1, \nu_2)$  distribution.



**FIGURE 9.16:** Two  $F$  distribution plots for different values for  $\nu_1$ , the first degree of freedom, but the same value for  $\nu_2$ . Note how the mode shifts to the right as  $\nu_1$  increases

### Applications

The  $F$  distribution is used extensively in statistical inference to make probability statements about the ratio of variances of random variables, providing the basis for the  $F$ -test. It is the theoretical probability tool for ANOVA (Analysis of Variance). Values of  $P(X \leq x)$  are traditionally tabulated for various values of  $\nu_1, \nu_2$  and selected values of  $x$ , and referred to as “ $F$ -tables.” Once again, computer programs have made such tables somewhat obsolete.

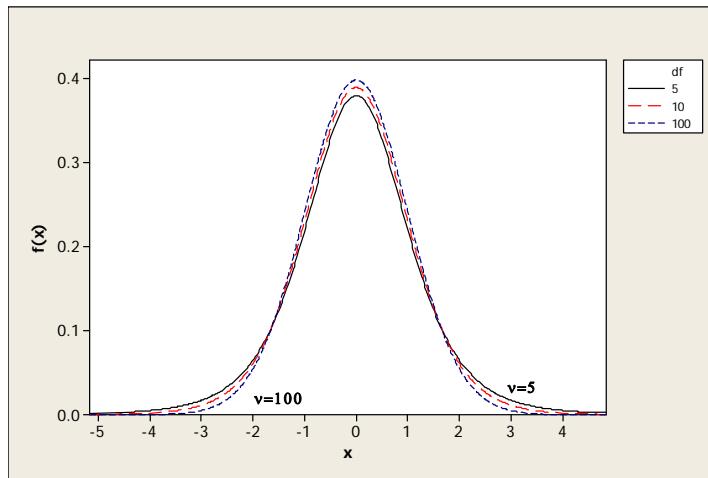
As shown in Part IV, the  $F$  distribution is one of the “central quartet” of pdfs at the core of statistical inference — the other three being the Gaussian (Normal)  $N(\mu, \sigma^2)$ , the Chi-square  $\chi^2(r)$ , and the Student  $t$ -distribution discussed next.

#### 9.3.5 Student’s t Random Variable

##### Basic Characteristics, Model and Remarks

Let  $Z$  be a standard normal random variable, (i.e.  $Z \sim N(0, 1)$ ); and let  $Y$ , a random variable independent of  $Z$ , possess a  $\chi^2(\nu)$  distribution. Then the random variable defined as the following ratio:

$$X = \frac{Z}{\sqrt{Y/\nu}} \quad (9.176)$$



**FIGURE 9.17:** Three  $t$ -distribution plots for degrees of freedom values  $\nu = 5, 10, 100$ . Note the symmetrical shape and the heavier tail for smaller values of  $\nu$ .

is a Student's  $t$  random variable  $t(\nu)$ . It can be shown that its pdf is:

$$f(x) = \frac{\Gamma\left[\frac{1}{2}(\nu + 1)\right]}{\sqrt{(\nu\pi)}\Gamma(\nu/2)} \frac{1}{\left(1 + \frac{x^2}{\nu}\right)^{(\nu+1)/2}}; -\infty < x < \infty \quad (9.177)$$

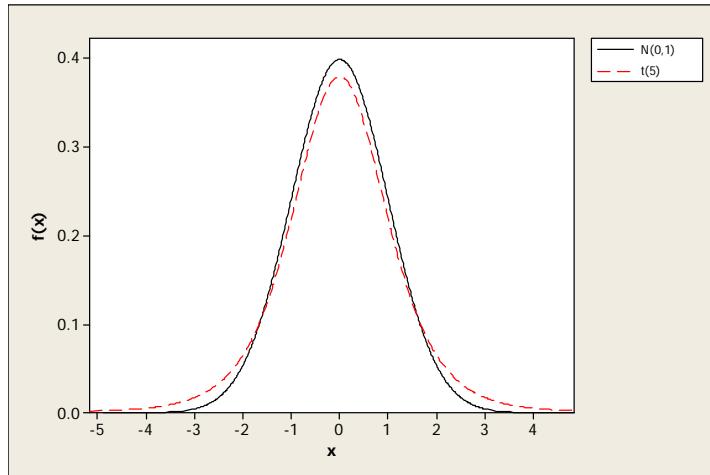
First derived in 1908 by W. S. Gossett (1876-1937), a Chemist at the Guinness Brewing Company who published under the pseudonym "Student", this  $f(x)$  is known as "Student's"  $t$ -distribution, or simply the  $t$ -distribution.

Even though it may not be immediately obvious from the somewhat awkward-looking form of  $f(x)$  shown above, the pdf for a  $t(\nu)$  random variable, is symmetrical about  $x = 0$  and appears like a heavier-tailed version of the standard normal  $N(0, 1)$  distribution. In fact, in the limit as  $\nu \rightarrow \infty$ , the  $t$ -distribution tends to  $N(0, 1)$ . In practice, for  $\nu \geq 50$ , the  $t$ -distribution is virtually indistinguishable from the standard normal distribution. These facts are illustrated in Figure 9.17, which shows  $t$ -distributions for three different degrees of freedom  $\nu = 5, 10, 100$ ; Figure 9.18, which compares the  $t(5)$  and standard normal distributions; and finally Figure 9.19, which shows that the  $t$ -distribution with 50 degrees of freedom is practically identical to the  $N(0, 1)$  distribution.

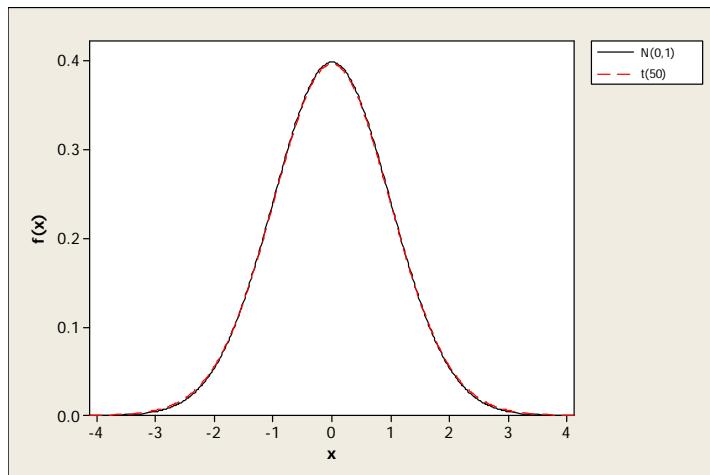
### Important Mathematical Characteristics

The following are some key mathematical characteristics of the  $t(\nu)$  random variable and its pdf:

1. **Characteristic parameter:**  $\nu$ , the degrees of freedom;



**FIGURE 9.18:** A comparison of the  $t$ -distribution with  $\nu = 5$  with the standard normal  $N(0, 1)$  distribution. Note the similarity as well as the  $t$ -distribution's comparatively heavier tail.



**FIGURE 9.19:** A comparison of the  $t$ -distribution with  $\nu = 50$  with the standard normal  $N(0, 1)$  distribution. The two distributions are practically indistinguishable.

2. **Mean:**  $E(X) = \mu = 0$ ; Mode = 0 = Median
3. **Variance:**  $Var(X) = \sigma^2 = \nu/(\nu - 2); \nu > 2$
4. **Higher Moments:** Coefficient of Skewness:  $\gamma_3 = 0$  for  $\nu > 3$  (indicating that the distribution is always symmetric);  
Coefficient of Kurtosis:

$$\gamma_4 = 3 \left( \frac{\nu - 2}{\nu - 4} \right); \nu > 4$$

This shows that for smaller values of  $\nu$ , the kurtosis exceeds the standard reference value of 3 for the normal random variable (implying a heavier *leptokurtic* tail); as  $\nu \rightarrow \infty$ , however,  $\gamma_4 \rightarrow 3$ .

Of the relationships between the  $t(\nu)$  random variable and other random variables, the most important are (i)  $\lim_{\nu \rightarrow \infty} t(\nu) \rightarrow N(0, 1)$ , as noted earlier; and (ii) If  $X \sim t(\nu)$ , then  $X^2 \sim F(1, \nu)$ .

### Applications

The  $t$ -distribution is used extensively in statistical inference, especially for comparing the means of two populations given only finite sample data. It is a key theoretical probability tool used in problems requiring tests of hypotheses involving means, providing the basis of the familiar “*t-test*”.

As with all the other distributions employed in data analysis and statistical inference, tables of tail area probabilities  $P(X \leq x)$  are available for various degrees of freedom values. Once again, the ability to compute these probabilities directly using computer programs is making such tables obsolete.

#### 9.3.6 The Cauchy Random Variable

##### Basic Characteristics, Model and Remarks

If  $Y_1$  is a Normal random variable with a  $N(0, \sigma_1^2)$  distribution, and  $Y_2$ , independent of  $Y_1$ , is a Normal random variable with a  $N(0, \sigma_2^2)$ , then the random variable,  $X$ , defined as the following ratio of the two random variables,

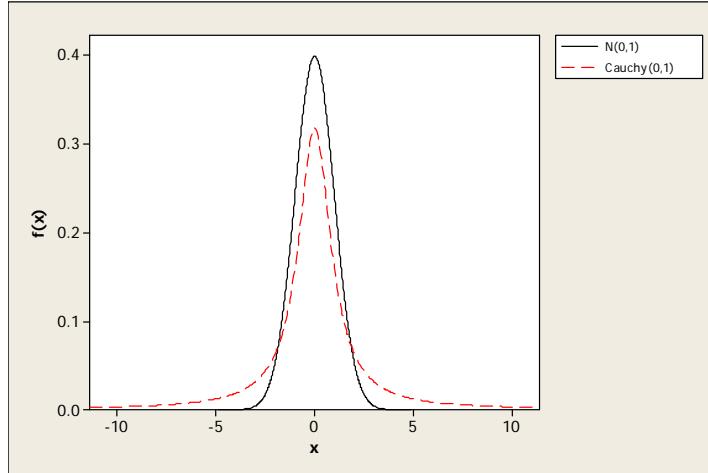
$$X = \frac{Y_1}{Y_2} \quad (9.178)$$

is a Cauchy random variable. If  $\sigma_1/\sigma_2$ , the ratio of the contributing variances is designated as  $\sigma$ , then it can be shown that the pdf for the Cauchy random variable is:

$$f(x) = \frac{1}{\pi\sigma} \frac{1}{\left(1 + \frac{x^2}{\sigma^2}\right)}; -\infty < x < \infty; \sigma > 0 \quad (9.179)$$

In particular, when  $\sigma_1 = \sigma_2 = 1$  (so that  $Y_1$  and  $Y_2$  are independent standard normal random variables),  $\sigma = 1$  above and the pdf becomes

$$f(x) = \frac{1}{\pi} \frac{1}{(1 + x^2)}; -\infty < x < \infty \quad (9.180)$$



**FIGURE 9.20:** A comparison of the standard Cauchy distributions with the standard normal  $N(0, 1)$  distribution. Note the general similarities as well as the Cauchy distribution's substantially heavier tail.

the expression for the standard Cauchy distribution — an expression that was derived in Example 6.9 in Chapter 6.

The pdf for the general Cauchy random variable,  $\mathcal{C}(\mu, \sigma^2)$ , is:

$$f(x) = \frac{1}{\pi\sigma} \frac{1}{\left[1 + \frac{(x-\mu)^2}{\sigma^2}\right]}; -\infty < x < \infty; \sigma > 0 \quad (9.181)$$

In a manner somewhat reminiscent of the  $t$ -distribution, the Cauchy distribution is also symmetric (about  $\mu$  in the general case, about 0 in the standard case), but with much heavier tails than the normal distribution. (See Fig 9.20.)

### Important Mathematical Characteristics

The Cauchy random variable is quite unusual in that its pdf has no finite moments, i.e.  $E(X^k)$  does not exist for any  $k > 0$ . The characteristic parameters  $\mu$  and  $\sigma$  are therefore *not* what one would expect—they *do not* represent the mean and standard deviation.

1. **Characteristic parameters:**  $\mu$ , is the location parameter;  $\sigma$ , is the scale parameter.
2. **Mean:**  $E(X)$  does not exist; Mode =  $\mu$  = Median
3. **Variance:**  $Var(X)$  does not exist; neither does any other moment.

Of the relationships between the Cauchy random variable and other random variables, the most notable are (i) By its very composition as a ratio of zero mean Gaussian random variables, if  $X$  is a Cauchy random variable, its reciprocal  $1/X$  is also a Cauchy random variable; (ii) The standard Cauchy random variable  $\mathcal{C}(0, 1)$  is a special (pathological) case of the  $t$  distribution with  $\nu = 1$  degrees of freedom. When we discuss the statistical implication of “degrees of freedom” in Part IV, it will be come clearer why  $\nu = 1$  is a pathological case.

### Applications

The Cauchy distribution is used mostly to represent otherwise symmetric phenomena where the occurrences of extreme values that are significantly far from the central values are not so rare. The most common application is in modeling high-resolution rates of price fluctuations in financial markets. Such data tend to have “heavier tails” and are hence poorly modeled by Gaussian distributions.

It is not difficult to see, from the genesis of the Cauchy random variable as a ratio of two Gaussians, why such applications are structurally reasonable. Price fluctuation rates are approximated as a ratio of  $\Delta P$ , the change in the price of a unit of goods, and  $\Delta t$  the time interval over which the price change has been computed. Both are independent random variables (prices may remain steady for a long time and then change rapidly over short periods of time) that tend, under normal elastic market conditions to fluctuate around some mean value. Hence,  $\Delta P/\Delta t$  will appear as a ratio of two Gaussian random variables and will naturally follow a Cauchy distribution.

### 9.4 Summary and Conclusions

Using precisely the same techniques and principles employed in the previous chapter, we have turned our attention in this chapter to the complementary task of model development for continuous random variables, with the discrete Poisson random variable serving as the connecting bridge between the two chapters. Our emphasis has been on the fundamentals of *how* each probability model arises for these continuous random variables, with the derivation details presented explicitly in many cases, or else left as exercises.

By design, we have encountered these continuous random variables and their models in family groups whose members share certain common structural traits: first, the Gamma family of strictly positive random variables, typically used to represent phenomena involving intervals of time and space (length, area, volume), or, as with the Chi-square random variable, squared and other variance-related phenomena; next the Gaussian family, with functional forms indicative of squared deviations from a target; and finally the

Ratio family of random variables strategically composed from ratios of *other* random variables. From the description of the phenomenon in question, the ensuing model derivation, the resulting model and its mathematical characterization, and from the illustrative examples given for each random variable, it should be clear for which practical phenomena these models are most applicable. And for some of these random variables, what we have seen in this chapter is no more than just a brief “cameo appearance;” we will most definitely see them and their models again, in their more natural application settings. In particular, when the topic of statistics and statistical inference is studied in detail in Part IV, we will be reacquainted with a quartet of random variables and pdfs whose role in such studies is dominant and central: the Gaussian distribution, for representing—precisely or approximately—an assortment of random variations related to experimental data and functions of such data; the Chi-square, Student’s  $t$ , and Fisher’s  $F$  distributions, for testing various hypotheses. Also, it should not be surprising when members of the Gamma family, especially the exponential and Weibull random variables, reappear to play central roles in the discussion of reliability and life testing of Chapter 21.

This is an admittedly long chapter; and yet, for all its length and breadth of coverage, it is by no means exhaustive. Some continuous random variables were not discussed at all; and the introduction to mixture distributions (where the parameters of a probability model is itself a random variable, leading to a compounding of models) was all-too-brief, restricted only to the Poisson-gamma model, which happens to lead to the negative binomial distribution. Some of the omitted pdfs have been introduced as exercises at the end of the chapter, e.g., the double exponential (Laplace), inverse gamma, logistic, and Pareto distributions, in the category of continuous pdfs, and the Beta-Binomial mixture model. Nevertheless, in terms of intended *scope* of coverage—the variety of continuous (and some mixed) probability models that have been discussed—one can consider the task begun in Chapter 8 as having now been concluded in this chapter. However, in terms of model development techniques—the “how” of probability model development, as opposed to the models themselves—there is one more topic to consider, when available information on the random variable of interest is incomplete. This is next chapter’s focus.

Table 9.1, similar to Table 8.2 in Chapter 8, is a summary of the main characteristics, models, and other important features of the continuous random variables discussed in this chapter. Fig 9.21 provides a schematic consolidation of *all* the random variables we have encountered, discrete and continuous, and the connections among them.

## REVIEW QUESTIONS

1. What are the four members of the Gamma family of random variables discussed in this chapter?
2. What are the common structural characteristics shared by the members of the

TABLE 9.1: Summary of probability models for continuous random variables

| Random Variable                           | Range               | Probability Model<br>$f(x)$   | Mean ( $\mu$ )<br>$E(X)$      | Variance ( $\sigma^2$ )<br>$Var(X)$                    | Relation to Other Variables   |
|---|---------------------|---|-------------------------------|--|---|
| Exponential<br>$\mathcal{E}(\beta)$       | $(0, \infty)$       | $\frac{1}{\beta}e^{-x/\beta}$   | $\beta$                       | $\beta^2$  | Inter-Poisson intervals   |
| Gamma                                     | $(0, \infty)$       | $\frac{1}{\beta^\alpha \Gamma(\alpha)} e^{-x/\beta} x^{\alpha-1}$                       | $\alpha\beta$                 | $\alpha\beta^2$  | $X_i \sim \mathcal{E}(\beta)$<br>$\sum_{i=1}^n X_i \sim \gamma(\alpha, \beta)$  |
| $\gamma(\alpha, \beta)$                   | $(0, \infty)$       | $\frac{1}{2^{r/2}\Gamma(r/2)} e^{-x/2} x^{r/2-1}$                                       | $r$                           | $2r$   | $\chi^2(r) = \gamma(r/2, 2)$<br>$X_i \sim N(0, 1) \Rightarrow$<br>$\sum_{i=1}^n X_i \sim \chi^2(n)$   |
| Chi-Square<br>$\chi^2(r)$                 | $(0, \infty)$       |   |                               |  |   |
| Weibull<br>$W(\zeta, \beta)$              | $(0, \infty)$       | $\frac{\zeta}{\beta} (x/\beta)^{\zeta-1} e^{-(x/\beta)^\zeta}$                          | $\beta\Gamma(1+1/\zeta)$      | $\beta^2\Gamma(1+2/\zeta) - \mu^2$                     | $Y \sim \mathcal{E}(\beta) \Rightarrow$<br>$Y^\zeta \sim W(\zeta, \beta)$   |
| Gaussian<br>Normal $N(\mu, \sigma^2)$     | $(-\infty, \infty)$ | $\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$           | $\mu$                         | $\sigma^2$   | $\lim_{n \rightarrow \infty} Bi(n, p)$  |
| Lognormal<br>$\mathcal{L}(\alpha, \beta)$ | $(0, \infty)$       | $\frac{1}{x\beta\sqrt{2\pi}} \exp\left\{-\frac{(\ln x - \alpha)^2}{2\beta^2}\right\}$   | $\exp(\alpha + \beta^2/2)$    | $\exp(2\alpha + \beta^2) \times (\exp(\beta^2) - 1)$   | $Y \sim N(\alpha, \beta^2) \Rightarrow$<br>$X = e^Y \sim \mathcal{L}(\alpha, \beta)$  |
| Rayleigh<br>$\mathcal{R}(b^2)$            | $(0, \infty)$       | $\frac{x}{b^2} \exp\left\{-\frac{x^2}{2b^2}\right\}$                                    | $b\sqrt{\pi/2}$               | $b^2(2 - \pi/2)$                                       | $Y_1, Y_2 \sim N(0, b^2) \Rightarrow$<br>$\sqrt{Y_1^2 + Y_2^2} \sim \mathcal{R}(b^2)$   |
| Beta<br>$B(\alpha, \beta)$                | $(0, 1)$            | $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$ | $\frac{\alpha}{\alpha+\beta}$ | $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ | $X_1 \sim \gamma(\alpha, 1); X_2 \sim \gamma(\beta, 1)$<br>$\frac{X_1}{X_1+X_2} \sim B(\alpha, \beta)$  |
| Uniform $U(a, b)$                         | $(a, b)$            | $\frac{1}{b-a}$   | $\frac{a+b}{2}$               | $\frac{(b-a)^2}{12}$                                   | $B(\alpha = \beta) = U(0, 1)$   |
| Fisher                                    | $(0, \infty)$       | See Eq (9.172)  | $\frac{\nu_2}{\nu_2-2}$       | See Eq (9.173)   | $Y_1 \sim \xi^2(\nu_1); Y_2 \xi^2(\nu_2) \Rightarrow$<br>$\frac{Y_1/\nu_1}{Y_2/\nu_2} \sim F(\nu_1, \nu_2)$   |
| $F(\nu_1, \nu_2)$                         |                     |   |                               |  |   |
| Student's $t(\nu)$                        | $(-\infty, \infty)$ | See Eq (9.177)  | 0                             | $\frac{\nu}{\nu-2}$                                    | $\lim_{\nu \rightarrow \infty} t(\nu) = N(0, 1)$<br>$Z \sim N(0, 1); Y \sim \chi^2(\nu) \Rightarrow$<br>$\frac{\sqrt{Y/\nu}}{\sqrt{Z/\nu}} \sim t(\nu)$ |
| Cauchy<br>$\mathcal{C}(0, 1)$             | $(-\infty, \infty)$ | $\frac{1}{\pi} \frac{1}{(1+x^2)}$   | Median=0                      | N/A  | $Y_1, Y_2, \sim N(0, 1) \Rightarrow$<br>$Y_1/Y_2 \sim \mathcal{C}(0, 1)$  |

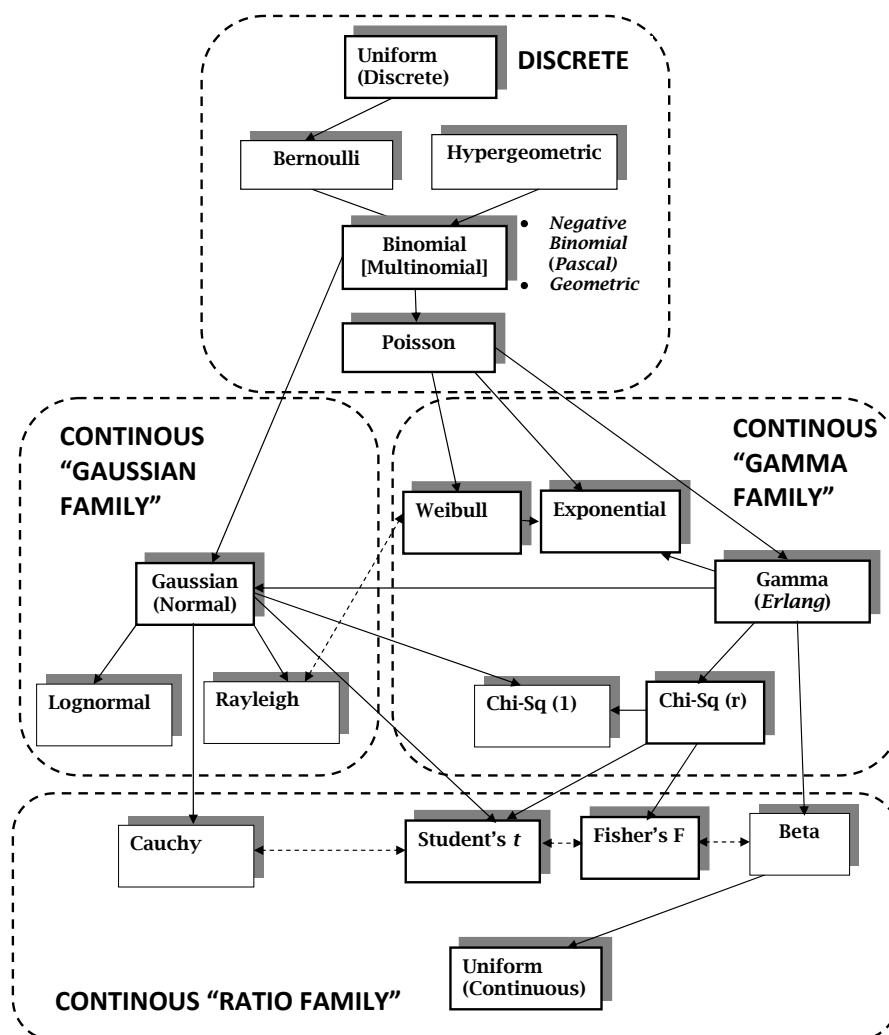


FIGURE 9.21: Common probability distributions and connections among them

Gamma family of random variables?

- 3.** Which member of the Gamma family of distributions is structurally different from the other three, and in what way is it different?
- 4.** What is the relationship between the exponential and the Poisson random variables?
- 5.** What are the basic characteristics of the exponential random variable?
- 6.** What is the probability model for the exponential random variable?
- 7.** How are the geometric and exponential random variables related?
- 8.** The exponential pdf finds application in what class of problems?
- 9.** Why is the exponential pdf known as a “memoryless” distribution? Are there other distributions with this characteristic?
- 10.** What are the basic characteristics of the gamma random variable?
- 11.** How is the gamma random variable related to the Poisson random variable?
- 12.** How are the gamma and exponential random variables related?
- 13.** What is the probability model for the gamma random variable and what do the parameters represent?
- 14.** Under what condition is the gamma distribution known as the Erlang distribution?
- 15.** What does it mean that the gamma random variable possesses a “reproductive” property?
- 16.** The gamma pdf finds application in what class of problems?
- 17.** What is the relationship between the Chi-square and gamma random variables?
- 18.** What are the basic characteristics of the Chi-square random variable?
- 19.** What is the probability model for the Chi-square random variable and what is the single parameter called?
- 20.** In what broad area does the Chi-square pdf find application?
- 21.** What differentiates the Weibull random variable from the exponential random variable?

- 22.** What are the basic characteristics of the Weibull random variable?
- 23.** What is the probability model for the Weibull random variable?
- 24.** Why is the Weibull pdf parameter  $\beta$  known as the “characteristic life”?
- 25.** The Weibull pdf finds application in what class of problems?
- 26.** What mixture pdf arises from a Poisson pdf whose parameter  $\lambda$  is gamma distributed?
- 27.** What are the three members of the Gaussian family of random variables discussed in this chapter?
- 28.** What are the common structural characteristics shared by the members of the Gaussian family of random variables?
- 29.** What are the three approaches used in this chapter to derive the probability model for the Gaussian random variable?
- 30.** What are the basic characteristics of the Gaussian random variable?
- 31.** What is the probability model for the Gaussian random variable and what do the parameters represent?
- 32.** In what broad area does the Gaussian pdf play an important role?
- 33.** What is the probability model for the standard normal random variable? How is the standard normal random variable related to the Gaussian random variable?
- 34.** What is the *z-score* of any particular value  $x_i$  of the general Gaussian random variable with mean  $\mu$ , and variance  $\sigma^2$ ? How is it useful for computing probabilities for general Gaussian distributions?
- 35.** What are the basic characteristics of the lognormal random variable?
- 36.** How is the lognormal random variable related to the Gaussian (normal) random variable?
- 37.** What is the probability model for the lognormal random variable?
- 38.** What trap is to be avoided in interpreting what the parameters of the lognormal distribution represent?
- 39.** What is the difference between the parameter  $\mu$  for the normal distribution and the corresponding parameter  $\alpha$  for the lognormal distribution in terms of how changes in each parameter modify the distribution it characterizes?

- 40.** Which phenomena are well-modeled by the lognormal pdf?
- 41.** What are the basic characteristics of the Rayleigh random variable?
- 42.** What is the probability model for the Rayleigh random variable?
- 43.** What is the relationship between the Weibull distribution and the Rayleigh distribution and why does this appear to be an odd coincidence?
- 44.** What are some rational justifications for why the Rayleigh random variable is related to the Weibull random variable?
- 45.** In what class of problems does the Rayleigh pdf find application?
- 46.** What are the four members of the Ratio family of random variables discussed in this chapter?
- 47.** What are the common structural characteristics shared by the members of the Ratio family of random variables?
- 48.** The Beta random variable is composed as a ratio of which random variables?
- 49.** What is the probability model for the Beta random variable?
- 50.** What are the various shapes possible with a Beta pdf, and what specific combinations of distribution parameters result in which shape?
- 51.** The Beta pdf provides a good model for what types of random phenomena?
- 52.** What is an inverted Beta random variable and how is it related to the Beta random variable?
- 53.** What are the basic characteristics of the (continuous) uniform random variable?
- 54.** What is the probability model for the (continuous) uniform random variable?
- 55.** How is the (continuous) uniform random variable related to the Beta random variable?
- 56.** What is the (continuous) uniform pdf used for mostly?
- 57.** Fisher's  $F$  random variable is composed as a ratio of which random variables?
- 58.** What is the relationship between Fisher's  $F$  distribution and the Beta distribution?
- 59.** What is the  $F$  distribution used for most extensively?

- 60.** What are the four central pdfs used most extensively in statistical inference?
- 61.** Student's  $t$  random variable is composed as a ratio of which random variables?
- 62.** What is the relationship between Student's  $t$  distribution and the standard normal distribution?
- 63.** What is the  $t$  distribution used for most extensively?
- 64.** The Cauchy random variable is composed as a ratio of which random variables?
- 65.** What is the probability model for the Cauchy random variable?
- 66.** How many moments exist for the Cauchy distribution?
- 67.** The Cauchy distribution is used mostly for what?

## EXERCISES

### Section 9.1

- 9.1** (i) On the same graph, plot the pdf for the discrete geometric  $G(0.25)$  and the continuous exponential  $\mathcal{E}(4)$  distributions. Repeat this for the following additional pairs of distributions:  $G(0.8)$  and  $\mathcal{E}(1.25)$ ; and  $G(0.5)$  and  $\mathcal{E}(2)$ .  
(ii) These plots show *specific* cases in which the pdf of the geometric random variable  $G(p)$  is seen to be a discretized version of the continuous pdf for the exponential random variable  $\mathcal{E}(1/p)$ , and vice-versa: that the  $\mathcal{E}(\beta)$  distribution is a continuous version of the discrete  $G(1/\beta)$  distribution. First, show that for the geometric random variable, the following relationship holds:

$$\frac{f(x+1) - f(x)}{f(x)} = p$$

which is a finite difference discretization of the expression,

$$\frac{df(x)}{f(x)} = p.$$

From here, establish the general result that the pdf of a geometric random variable  $G(p)$  is the discretized version of the pdf of the exponential random variable,  $\mathcal{E}(1/p)$ .

- 9.2** Establish that the median of the exponential random variable,  $\mathcal{E}(\beta)$ , is  $\beta \ln 2$ , and that its hazard function is

$$h(t) = \frac{1}{\beta} = \eta$$

- 9.3** Given two independent random variables,  $X_1$  and  $X_2$ , with identical exponential  $\mathcal{E}(\beta)$  distributions, show that the pdf of their difference,

$$Y = X_1 - X_2 \quad (9.182)$$

is the double exponential (or Laplace) distribution defined as:

$$f(y) = \frac{\eta}{2} e^{-\eta|y|}; -\infty < y < \infty \quad (9.183)$$

where  $\eta = 1/\beta$ .

**9.4** Revisit Exercise 9.3. Directly from the pdf in Eq (15.183), and the formal definitions of moments of a random variable, obtain the mean and variance of  $Y$ . Next, obtain the mean and variance from Eq (15.182) by using the result that because the two random variables are independent,

$$\begin{aligned} E(Y) &= E(X_1) - E(X_2) \\ Var(Y) &= Var(X_1) + Var(X_2) \end{aligned}$$

**9.5** Given that  $X \sim \mathcal{E}(1)$ , i.e., an exponentially distributed random variable with parameter  $\beta = 1$ , determine the following probabilities:

- (i)  $P(X - \mu_X \geq 3\sigma_X)$  where  $\mu_X$  is the mean value, and  $\sigma_X$  is the standard deviation, the positive square root of the variance,  $\sigma^2$ .
- (ii)  $P(\mu_X - 2\sigma_X < X < \mu_X + 2\sigma_X)$

**9.6** (i) Establish the result in Eq (9.17).

(ii) From the definition of the gamma function:

$$\Gamma(\alpha) = \int_0^\infty e^{-y} y^{\alpha-1} dy \quad (9.184)$$

establish the following properties:

- (a)  $\Gamma(1) = 1$ ;
- (b)  $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$ ;  $\alpha > 1$ ;
- (c)  $\Gamma(\alpha) = (\alpha - 1)!$  for integer  $\alpha$ .

**9.7** For the random variable  $X$  with the  $\gamma(\alpha, \beta)$  pdf:

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} e^{-x/\beta} x^{\alpha-1}; x > 0; \alpha, \beta > 0$$

establish the following about the moments of this pdf:

- (i)  $\mu = \alpha\beta$ ;
- (ii)  $\sigma^2 = \alpha\beta^2$ ;
- (iii)  $M(t) = (1 - \beta t)^{-\alpha}$

**9.8** Show that if  $Y_i$  is an exponential random variable with parameter  $\beta$ , i.e.  $Y_i \sim \mathcal{E}(\beta)$ , then the random variable  $X$  defined as in Eq (9.37), i.e.,:

$$X = \sum_{i=1}^{\alpha} Y_i$$

is the gamma random variable  $\gamma(\alpha, \beta)$ .

**9.9** Establish the following results for the gamma random variable:

- (i) If  $X_i, i = 1, 2, \dots, n$ , are  $n$  independent gamma random variables, each with different shape parameters  $\alpha_i$  but a common scale parameter  $\beta$ , i.e.  $X_i \sim \gamma(\alpha_i, \beta)$ , then the random variable  $Y$  defined as:

$$Y = \sum_{i=1}^n X_i$$

is also a gamma random variable, with shape parameter  $\alpha^* = \sum_{i=1}^n \alpha_i$  and scale parameter  $\beta$ , i.e.  $Y \sim \gamma(\alpha^*, \beta)$ .

- (ii) Show that the random variable  $Z$  defined as

$$Z = c \sum_{i=1}^n X_i$$

where  $c$  is a constant, is also a gamma random variable with shape parameter  $\alpha^* = \sum_{i=1}^n \alpha_i$  but with scale parameter  $c\beta$ , i.e.  $Z \sim \gamma(\alpha^*, c\beta)$ .

**9.10** The distribution of residence times in a standard size continuous stirred tank reactor (CSTR) is known to be exponential with  $\beta = 1$ , i.e.,  $\mathcal{E}(1)$ . If  $X$  is the residence time for a reactor that is five times the standard size, then its distribution is also known as  $\mathcal{E}(0.2)$ . On the other hand,  $Y$ , the residence time in an ensemble of five identical, standard size CSTR's in series, is known to be gamma distributed with  $\alpha = 5; \beta = 1$ .

(i) Plot the pdf  $f(x)$  for the single large CSTR's residence time distribution and the pdf  $f(y)$  for the ensemble of five identical small reactors in series. Determine the mean residence time in each case.

(ii) Compute  $P(Y \leq 5)$  and compare with  $P(X \leq 5)$

**9.11** Given that  $X \sim \gamma(\alpha, \beta)$ , show that the pdf for  $Y$ , the Inverse Gamma, IG, random variable defined by  $Y = 1/X$  is given by:

$$f(y) = \frac{(1/\beta)^\alpha}{\Gamma(\alpha)} e^{-(1/\beta)/y} y^{-\alpha-1}; 0 < y < \infty \quad (9.185)$$

Determine the mean, mode and variance for this random variable.

**9.12** Establish the following results that (i) if  $Y \sim \mathcal{E}(\beta)$ , then

$$X = Y^{1/\zeta}$$

is a  $W(\zeta, \beta)$  random variable; and (ii) conversely, that if  $X \sim W(\zeta, \beta)$  then

$$Y = X^\zeta$$

is an  $\mathcal{E}(\beta)$  random variable.

**9.13** A fluidized bed reactor through which chlorine gas flows has a temperature probe that fails periodically due to corrosion. The length of time (in days) during which the temperature probe functions is known to be a Weibull distributed random variable  $X$ , with parameters  $\beta = 10; \zeta = 2$ .

- (i) Determine the number of days each probe is expected to last.

- (ii) If the reactor operator wishes to run a product campaign that lasts continuously for 20 days, determine the probability of running this campaign without having to replace the temperature probe.
- (iii) What is the probability that the probe will function for anywhere between 10 and 15 days?
- (iv) What is the probability that the probe will fail on or before the 10<sup>th</sup> day?

**9.14** Suppose that the time-to-failure (in minutes) of certain electronic device components, when subjected to continuous vibrations, may be considered as a random variable having a Weibull( $\zeta, \beta$ ) distribution with  $\zeta = 1/2$  and  $\beta^* = \beta^{-\zeta} = 1/10$ : first find how long we may expect such a component to last, and then find the probability that such a component will fail in less than 5 hours.

### Section 9.2

**9.15** Given two random variables  $X$  and  $Z$  related according to Eq (9.78), i.e.,

$$Z = \frac{X - \mu_x}{\sigma_x},$$

where, as defined in the text,  $E(X) = \mu_x$  and  $Var(X) = \sigma_x^2$ ,

- (i) Show that  $E(Z) = 0$  and  $Var(Z) = 1$ .
- (ii) If the pdf of  $Z$  is as given in Eq (9.90), i.e.,

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

determine the pdf for the random variable  $X$  and hence confirm Eq (9.91).

**9.16** Given a Gaussian distributed random variable,  $X$ , with  $\mu = 100$ ;  $\sigma = 10$ , determine the following probabilities:

- (i)  $P(-1.96\sigma < X - \mu < 1.96\sigma)$  and  $P(-3\sigma < X - \mu < 3\sigma)$
- (ii)  $P(X > 123)$  and  $P(74.2 < X < 126)$

**9.17** Given  $Z$ , a standard normal random variable, determine the specific variate  $z_0$  that satisfies each of the following probability statements:

- (i)  $P(Z \geq z_0) = 0.05$ ;  $P(Z \geq z_0) = 0.025$
- (ii)  $P(Z \leq z_0) = 0.025$ ;  $P(Z \geq z_0) = 0.10$ ;  $P(Z \leq z_0) = 0.10$
- (iii)  $P(|Z| \leq z_0) = 0.00135$

**9.18** Given  $Z$ , a standard normal random variable, determine the following probabilities

- (i)  $P(-1.96 < Z < 1.96)$  and  $P(-1.645 < Z < 1.645)$
- (ii)  $P(-2 < Z < 2)$  and  $P(-3 < Z < 3)$
- (iii)  $P(|Z| \leq 1)$

**9.19** Consider the random variable  $X$  with the following pdf:

$$f(x) = \frac{e^{-x}}{(1 - e^{-x})^2} \quad (9.186)$$

This is the pdf of the (standard) logistic distribution.

- (i) Show that  $E(X) = 0$  for this random variable.  
(ii) On the same graph, plot this pdf and the standard normal pdf. Compare and contrast the two pdfs. Discuss under which condition you would recommend using the logistic distribution instead of the standard normal distribution.

**9.20** Let  $X_i; i = 1, 2, \dots, n$ , be  $n$  independent Gaussian random variables with identical distributions  $N(\mu, \sigma^2)$ ; show that the random variable defined as in Eq (9.138), i.e.,

$$Y = \sum_{i=1}^n \left( \frac{X_i - \mu}{\sigma} \right)^2$$

possesses a  $\chi^2(n)$  distribution.

**9.21** Show that if the random variable  $Y$  has a normal  $N(\mu, \sigma^2)$  distribution, then the random variable  $X$  defined as

$$X = e^Y$$

has a lognormal distribution, with parameters  $\alpha$  and  $\beta$ , as shown in Eq (9.143). Obtain an explicit relationship between  $(\alpha, \beta)$  and  $(\mu, \sigma^2)$ .

**9.22** Revisit Exercise 9.21 and establish the reciprocal result that if the random variable  $X$  has a lognormal  $\mathcal{L}(\alpha, \beta)$ , then the random variable  $Y$  defined as

$$Y = \ln X$$

has a normal  $N(\mu, \sigma^2)$  distribution.

**9.23** Given a lognormal distributed random variable  $X$  with parameters  $\alpha = 0; \beta = 0.2$ , determine its mean,  $\mu_X$ , and variance,  $\sigma_X^2$ ; on the same graph, plot the pdf,  $f(x)$ , and that for the Gaussian random variable with the same mean and variance as  $X$ . Compare the two plots.

**9.24** Revisit Exercise 9.23. Compute  $P(\mu_X - 1.96\sigma_X < X < \mu_X + 1.96\sigma_X)$  from a lognormal distribution. Had this random variable been mistakenly assumed to be Gaussian with the same mean and variance, compute the same probability and compare the results.

**9.25** Show that if  $Y_1 \sim N(0, b^2)$  and  $Y_2 \sim N(0, b^2)$ , then

$$X = \sqrt{Y_1^2 + Y_2^2} \quad (9.187)$$

possesses a Rayleigh  $\mathcal{R}(b)$  distribution, with pdf given by Eq (9.153).

**9.26** Given a random variable  $X$  with a Rayleigh  $\mathcal{R}(b)$  distribution, obtain the pdf of the random variable  $Y$  defined as

$$Y = X^2$$

### Section 9.3

**9.27** Confirm that if a random variable  $X$  has a Beta  $B(\alpha, \beta)$  distribution, the mode of the pdf occurs at:

$$x^* = \frac{\alpha - 1}{\alpha + \beta - 2} \quad (9.188)$$

and hence deduce that (a) no mode exists when  $0 < \alpha < 1$ , and  $\alpha + \beta > 2$ ; and (b) that when a mode exists, this mode and the mean will coincide if, and only if,  $\alpha = \beta$ . (*You may simply recall the expression for the mean given in the text; you need not rederive it.*)

**9.28** The Beta-Binomial mixture distribution arises from a Binomial  $Bi(n, p)$  random variable,  $X$ , whose parameter  $p$ , rather than being constant, has a Beta distribution, i.e., it consists of a conditional distribution,

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x}$$

in conjunction with the marginal distribution for  $p$ ,

$$f(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1}; 0 < p < 1; \alpha > 0; \beta > 0$$

Obtain the expression for  $f(x)$ , the resulting Beta-Binomial pdf.

**9.29** Given a random variable  $X$  with a standard uniform  $U(0, 1)$  distribution, i.e.,

$$f(x) = 1; 0 < x < 1$$

and any two points  $a_1, a_2$  in the interval  $(0, 1)$ , such that  $a_1 < a_2$  and  $a_1 + a_2 \leq 1$ , show that

$$P[a_1 < X < (a_1 + a_2)] = a_2.$$

In general, if  $f(x)$  is uniform in the interval  $(a, b)$ , and if  $a \leq a_1$ ,  $a_1 < a_2$  and  $a_1 + a_2 \leq b$ , show that:

$$P[a_1 < X < (a_1 + a_2)] = \frac{a_2}{(b-a)}$$

**9.30** For a random variable  $X$  that is uniformly distributed over the interval  $(a, b)$ :

- (i) Determine  $P(X > [\omega a + (1-\omega)b])$ ;  $0 < \omega < 1$ ;
- (ii) For the specific case where  $a = 1, b = 3$ , determine  $P(\mu_X - 2\sigma_X < X < \mu_X + 2\sigma_X)$  where  $\mu_X$  is the mean of the random variable, and  $\sigma_X$  is the positive square root of its variance.
- (iii) Again for the specific case where  $a = 1, b = 3$ , find the symmetric interval  $(\xi_1, \xi_2)$  around the mean,  $\mu_X$ , such that  $P(\xi_1 < X < \xi_2) = 0.95$

**9.31** Consider the random variable,  $X$ , with pdf:

$$f(x) = (\alpha - 1)x^{-\alpha}; x \geq 1; \alpha > 2 \quad (9.189)$$

known as a Pareto random variable.

- (i) Show that for this random variable,

$$E(X) = \frac{\alpha - 1}{\alpha - 2}; \alpha > 2$$

(ii) Determine the median and the variance of  $X$ .

**9.32** Given a random variable  $X$  that has an  $F(49, 49)$  distribution, determine the following probabilities:

- (i)  $P(X \geq 1)$
- (ii)  $P(X \geq 2); P(X \leq 0.5)$
- (iii)  $P(X \geq 1.76); P(X \leq 0.568)$

**9.33** Given a random variable  $X$  that has an  $F(\nu_1, \nu_2)$  distribution, determine the value  $x_0$  such that (a)  $P(X \geq x_0) = 0.025$ ; (b)  $P(X \leq x_0) = 0.025$  for the following specific cases:

- (i)  $\nu_1 = \nu_2 = 49$
- (ii)  $\nu_1 = \nu_2 = 39$
- (iii)  $\nu_1 = \nu_2 = 29$

(iv) Comment on the effect that reducing the degrees of freedom has on the various values of  $x_0$ .

**9.34** Given a random variable  $X$  that has a  $t(\nu)$  distribution, determine the value of  $x_0$  such that  $P(|X| < x_0) = 0.025$  for (i)  $\nu = 5$ ; (ii) (i)  $\nu = 25$ ; (iii)  $\nu = 50$ ; (iv)  $\nu = 100$ . Compare your results with the single value of  $x_0$  such that  $P(|X| < x_0) = 0.025$  for a standard normal random variable  $X$ .

**9.35** Plot on the same graph the pdfs for a Cauchy  $\mathcal{C}(5, 4)$  random variable and for a Gaussian  $N(5, 4)$  random variable. Compute the probability  $P(X \geq \mu + 1.96\sigma)$  for each random variable, where, for the Gaussian random variable,  $\mu$  and  $\sigma$  are the mean and standard deviation (positive square root of the variance) respectively; for the Cauchy distribution,  $\mu$  and  $\sigma$  are, the location and scale parameters, respectively.

**9.36** Plot on the same graph the pdf for the logistic distribution given in Eq (9.186) of Exercise 9.19 and that of the standard Cauchy random variable. Which pdf has the heavier tails?

## APPLICATION PROBLEMS

**9.37** The waiting time in days between the arrival of tornadoes in a county in south central United States is known to be an exponentially distributed random variable whose mean value remains constant throughout the year. Given that the probability is 0.6 that more than 30 days will elapse between tornadoes, determine the expected number of tornadoes in the next 90 days.

**9.38** The time-to-failure,  $T$ , of an electronic component is known to be an exponentially distributed random variable with pdf

$$f(t) = \begin{cases} \eta e^{-\eta t}; & 0 < x < \infty \\ 0; & \text{elsewhere} \end{cases} \quad (9.190)$$

where the “failure rate,”  $\eta = 0.075$  per 100 hours of operation.

(i) If the component “reliability function”  $R_i(t)$  is defined as

$$R_i(t) = P(T > t) \quad (9.191)$$

i.e., the probability that the component functions at least up until time  $t$ , obtain an explicit expression for  $R_i(t)$  for this electronic component.

(ii) A system consisting of two such components in *parallel* functions if at least one of them functions. Again assuming that both components are identical, find the system reliability  $R_p(t)$  and compute  $R_p(1000)$ , the probability that the system survives at least 1,000 hours of operation.

**9.39** Life-testing results on a first generation microprocessor-based (computer-controlled) toaster indicate that  $X$ , the life-span (in years) of the central control chip, is a random variable that is reasonably well-modeled by the exponential pdf:

$$f(x) = \eta e^{-\eta x}; x > 0 \quad (9.192)$$

with  $\eta = 0.16$ . A malfunctioning chip will have to be replaced to restore proper toaster operation.

- (i) The warranty for the chip is to be set at  $x_w$  years (in integers) such that no more than 15% would have to be replaced before the warranty period expires. Find  $x_w$ .
- (ii) In planning for the second generation toaster, design engineers wish to set a target value  $\eta = \eta_2^*$  to aim for such that 85% of the second generation chips survive beyond 3 years. Determine  $\eta_2^*$  and interpret your results in terms of the implied “fold increase” in mean life-span from the first to the second generation of chips.

**9.40** The table below shows frequency data on distances between DNA replication origins (inter-origin distances), measured *in vivo* in Chinese Hamster Ovary (CHO) cells by Li *et al.*, (2003)<sup>7</sup>, as reported in Chapter 7 of Birtwistle (2008)<sup>8</sup>. The data is similar to that in Fig 9.3 in the text.

| Inter-Origin Distance (kb) | Relative Frequency<br>$f_r(x)$ |
|----------------------------|--------------------------------|
| 0                          | 0.00                           |
| 15                         | 0.02                           |
| 30                         | 0.20                           |
| 45                         | 0.32                           |
| 60                         | 0.16                           |
| 75                         | 0.08                           |
| 90                         | 0.11                           |
| 105                        | 0.03                           |
| 120                        | 0.02                           |
| 135                        | 0.01                           |
| 150                        | 0.00                           |
| 165                        | 0.01                           |

- (i) Determine the mean (average) and variance of the CHO cells inter-origin distance.
- (ii) If this is a gamma distributed random variable, use the results in (i) to provide reasonable values for the gamma distribution parameters. On the same graph, plot the frequency data and the gamma model fit. Comment on the model fit to the data.

<sup>7</sup>Li, F., Chen, J., Solessio, E. and Gilbert, D. M. (2003). “Spatial distribution and specification of mammalian replication origins during G1 phase.” *J Cell Biol* 161, 257-66.

<sup>8</sup>M. R. Birtwistle, (2008). *Modeling and Analysis of the ErbB Signaling Network: From Single Cells to Tumorigenesis*, PhD Dissertation, University of Delaware.

(iii) It is known that DNA synthesis is initiated at replication origins, which are distributed non-uniformly throughout the genome, at an average rate of  $r$  origins per kb. However, in some mammalian cells, because there is a non-zero probability that any particular replication origin will *not* fire, some potential origins are skipped over so that in effect,  $k$  of such “skips” must take place (on average) *before* DNA synthesis can begin. What do the values estimated for the gamma distribution imply about the physiological parameters  $r$  and  $k$ ?

**9.41** The storage time (in months) until a collection of long-life Li/SO<sub>4</sub> batteries become unusable was modeled in Morris (1987)<sup>9</sup> as a Weibull distributed random variable with  $\zeta = 2$  and  $\beta = 10$ . Let us refer to this variable as the battery’s “maximum storage life,” MSL.

- (i) What is the “most likely” value of the MSL? (By definition, the “most likely” value is that value of the random variable for which the pdf attains a maximum.)
- (ii) What is the median MSL? By how much does it differ from the expected MSL?
- (iii) What is the probability that a battery has an MSL value exceeding 18 months?

**9.42** A brilliant paediatrician has such excellent diagnostic skills that without resorting to expensive and sometimes painful tests, she rarely misdiagnoses what ails her patients. Her overall average misdiagnosis rate of 1 per 1,000 consultations is all the more remarkable given that many of her patients are often too young to describe their symptoms adequately—when they can describe them at all; the paediatrician must therefore often depend on indirect information extracted from the parents and guardians during clinic visits. Because of her other responsibilities in the clinic, she must limit her patient load to precisely 10 per day for 325 days a year. While the total number of her misdiagnoses is clearly a Poisson random variable, the Poisson parameter  $\lambda = \eta t$  is not constant because of the variability in her patient population, both in age and in the ability of parents and guardians to communicate effectively on behalf of their non-verbal charges. If  $\lambda$  has a gamma distribution with  $\alpha = 13$ ,  $\beta = 0.25$ ,

- (i) Determine the probability that the paediatrician records exactly 3 misdiagnoses in a year; determine also the probability of recording 3 or fewer misdiagnoses.
- (ii) Compare these probabilities with the corresponding ones computed using a standard Poisson model with a constant parameter.

**9.43** The year-end bonuses of cash, stock and stock options (in thousands of US dollars) given to senior technical researchers in a leading chemical manufacturing company, each year over a five-year period from 1990–1994, has a lognormal distribution with scale parameter  $\alpha = 3$  and shape parameter  $\beta = 0.5$ .

- (i) Determine the probability that someone selected at random from this group received a bonus of \$20,000 or higher.
- (ii) If a bonus of \$100,000 or higher is considered a “Jumbo” bonus, what percentage of senior technical researchers received such bonuses during this period?
- (iii) If a bonus in the range \$10,000–\$30,000 is considered more typical, what percentage received this typical bonus?

---

<sup>9</sup>Morris, M.D. (1987). “A sequential experimental design for estimating a scale parameter from quantal life testing data.” *Technometrics*, 29, 173–181

**9.44** The home prices (in thousands of dollars) in a county located in the upper mid-Atlantic region of the United States is a lognormal random variable with a median of 403 and a mode of 245.

- (i) What percentage of the homes in this region cost more than \$500,000?
- (ii) If a home is considered “affordable” in this region if it costs between \$150,000 and \$300,000, what percentage of homes fall into this category?
- (iii) Plot the pdf for this random variable. Compute its mean and indicate its value on the plot along with the value given for the median. Which seems to be more “representative” of the central location of the distribution, the mean or the median?

**9.45** If the proportion of students who obtain failing grades on a foreign University’s highly competitive annual entrance examination can be considered as a Beta  $B(2, 3)$  random variable,

- (i) What is the mode of this distribution, and what percentage of students can be expected to fail annually?
- (ii) Determine the probability that over 90% of the students will pass this examination in any given year.
- (iii) The proportion of students from an elite college preparatory school (located in this same foreign country) who fail this same entrance examination has a Beta  $B(1, 7)$  distribution. Determine the percentage of this select group of students who can be expected to fail; also, determine the probability that over 90% of these elite students will pass this examination in any given year.
- (iv) Do these results mean that the elite college preparatory school does better in getting its students admitted into this highly selective foreign University?

**9.46** The place kicker on a team in the American National Football League (NFL) has an all-time success rate (total number field goals “made” divided by total number of field goals attempted) of 0.82 on field goal attempts of 55 yards or shorter. An attempt to quantify his performance with a probability model resulted in a Beta  $B(4.5, 1)$  distribution.

- (i) Is this model consistent with the computed all-time success rate?
- (ii) To be considered an elite place kicker, the success rate from this distance ( $D \leq 55$ ) needs to improve to at least 0.9. Determine the probability that this particular place kicker achieves elite status in any season, assuming that he maintains his current performance level.
- (iii) It is known that the computed probability of attaining elite status is sensitive to the model parameters, especially  $\alpha$ . For the same fixed value  $\beta = 1$ , compute the probability of attaining elite status for the values  $\alpha = 3.5, 4.0, 4.5, 5.0, 5.5$ . Plot these probabilities as a function of  $\alpha$ .

**9.47** If the fluorescence signals obtained from a test spot and the reference spot on a microarray—a device used to quantify changes in gene expression—is represented as random variables  $X_1$  and  $X_2$  respectively, it is possible to show that if these variables can be assumed to be independent, then they are reasonably represented by gamma distributions. In this case, the “fold change” ratio

$$Y = \frac{X_1}{X_2} \quad (9.193)$$

indicative of the “fold increase” (or decrease) in the signal intensity between test

and reference conditions, has the inverted Beta distribution, with the pdf

$$f(y) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{y^{\alpha-1}}{(1+y)^{\alpha+\beta}}; y > 0; \alpha > 0; \beta > 0 \quad (9.194)$$

The theoretical distribution with parameters  $\alpha = 4.8$ ;  $\beta = 2.1$  fit the fold change ratio for a particular set of data. Because of the detection threshold of the measurement technology, the genes in question are declared to be *overexpressed* only if  $Y \geq 2$ ; if  $0.5 \leq Y < 2$ , the conclusion is that there is insufficient evidence of “differential expression.”

- (i) Determine the expected fold change ratio.
- (ii) Determine the probability that a gene selected at random from this population will be identified as overexpressed.
- (iii) Determine the probability that there will be insufficient evidence to conclude that there is differential expression. (*Hint:* it may be easier to invoke the result that the variable  $Z$ , defined as  $Z = X_1/(X_1 + X_2)$ , has a Beta distribution with the same values of  $\alpha$  and  $\beta$  given for the inverted Beta distribution. In this case, the probabilities can be computed in terms of  $Z$  rather than  $Y$ .)

**9.48** The sample variance for the yield data presented in Chapter 1 may be determined as  $s_A^2 = 2.05$  for process  $A$ , and  $s_B^2 = 7.62$  for process  $B$ . If, but for random variability in the data, these variances are the same, then the ratio

$$x_{AB} = \frac{s_A^2}{s_B^2}$$

should be approximately equal to 1, but will not be exactly so (because of random variability). Given that if these two variances are theoretically the same, then this ratio is known to be a random variable  $X$  with an  $F(49, 49)$  distribution,

- (i) Determine the probability  $P(X \geq x_{AB})$ , that this random variable takes a value as large as, or larger than, the computed  $x_{AB}$  by pure chance alone, when the two variances are in fact the same.
- (ii) Determine the values  $f_1$  and  $f_2$  such that

$$\begin{aligned} P(X \leq f_1) &= 0.025 \\ P(X \geq f_2) &= 0.025 \end{aligned}$$

- (iii) What do these results imply about the plausibility of the conjecture that but for random variability, the variances of the data obtained from the two processes are in fact the same?



# Chapter 10

---

## *Information, Entropy and Probability Models*

|                       |  |     |
|-----------------------|--|-----|
| 10.1                  | Uncertainty and Information .....                                | 336 |
| 10.1.1                | Basic Concepts .....   | 336 |
| 10.1.2                | Quantifying Information .....                                    | 337 |
| 10.2                  | Entropy .....  | 338 |
| 10.2.1                | Discrete Random Variables .....                                  | 338 |
| 10.2.2                | Continuous Random Variables .....                                | 340 |
| 10.3                  | Maximum Entropy Principles for Probability Modeling .....        | 344 |
| 10.4                  | Some Maximum Entropy Models .....                                | 344 |
| 10.4.1                | Discrete Random Variable; Known Range .....                      | 345 |
| 10.4.2                | Discrete Random Variable; Known Mean .....                       | 346 |
| 10.4.3                | Continuous Random Variable; Known Range .....                    | 347 |
| 10.4.4                | Continuous Random Variable; Known Mean .....                     | 348 |
| 10.4.5                | Continuous Random Variable; Known Mean and Variance .....        | 349 |
| 10.4.6                | Continuous Random Variable; Known Range, Mean and Variance ..... | 350 |
| 10.5                  | Maximum Entropy Models from General Expectations .....           | 351 |
| 10.5.1                | Single Expectations .....  | 351 |
| Discrete Case .....   | 351  |     |
| Continuous Case ..... | 352  |     |
| 10.5.2                | Multiple Expectations .....                                      | 352 |
| 10.6                  | Summary and Conclusions .....                                    | 354 |
|                       | REVIEW QUESTIONS .....   | 355 |
|                       | EXERCISES .....  | 357 |
|                       | APPLICATION PROBLEMS .....                                       | 360 |

*For since the fabric of the universe is most perfect  
and the work of a most wise Creator,  
nothing at all takes place in the universe in which  
some rule of maximum or minimum does not appear*

Leonhard Euler (1707–1783)

The defining characteristic of the random variable is that uncertainty in individual outcomes co-exists with regularity in the aggregate ensemble. This aggregate ensemble is conveniently characterized by the pdf,  $f(x)$ ; and, as shown in the preceding chapters, given all there is to know about the phenomenon behind the random variable,  $X$ , one can derive expressions for  $f(x)$  from first principles. There are many practical cases, however, where the available information is insufficient to specify the full pdf. One can still obtain reasonable probability models from such incomplete information, but this will require an alternate view of the outcomes of random experiments in terms of the *infor-*

*mation* each conveys, from which derives the concept of the “entropy” of a random variable. This chapter is concerned first with introducing the concept of *entropy* as a means of quantifying *uncertainty* in terms of the amount of information conveyed by the observation of a random variable’s outcome. We then subsequently present a procedure that utilizes entropy to specify full pdfs in the face of incomplete information. The chapter casts most of the results of the two previous chapters in a different context that will be of interest to engineers and scientists.

## 10.1 Uncertainty and Information

### 10.1.1 Basic Concepts

Consider that after performing an experiment, a discrete random variable  $X$  is observed to take on the specific value  $x_i$ —for example, after running an experimental fiber spinning machine for 24 hours straight,  $X$ , the total number of line breaks during this period, is found to be 3. It is of interest to ask:

How much *information* about  $X$  is conveyed in the result  $X = x_i$ ? i.e. How much does the observation that  $X = x_i$  add to our knowledge about the random variable  $X$ ?

Equivalently, we could ask: “Prior to making the observation, how much *uncertainty* is associated with predicting the outcome  $X = x_i$ ?” thereby recasting the original question conversely as:

How much *uncertainty* is resolved by the specific observation that  $X = x_i$ ?

Clearly the answer to either version of this question depends on how much inherent variability is associated with the random variable  $X$ . If  $P(X = x_i)$  is relatively high (making it more likely than not that the event  $X = x_i$  will occur), the actual observation will not be very informative in the complementary sense that (i) we could have deduced this outcome with a high degree of certainty *before* performing the experiment; and (ii) there was not much uncertainty to be resolved by this observation. In the extreme, the occurrence of a *certain* event is thus entirely uninformative since there was no uncertainty to begin with, and therefore the observation adds nothing to what we already knew before the experiment.

Conversely, if  $P(X = x_i)$  is relatively low, it is less likely that  $X$  will take the value  $x_i$ , and the actual observation  $X = x_i$  will be quite informative. The high degree of uncertainty associated with predicting *a-priori* the occurrence of a rare event makes the actual observation of such an event (upon experimentation) very informative.

Let us illustrate with the following example. Case 1 involves a bag containing exactly two balls, 1 red, 1 blue. For Case 2, we add to this bag 10 green, 10 black, 6 white, 6 yellow, 3 purple and 3 orange balls to bring the total to 40 balls. The experiment is to draw a ball from this bag and to consider for each case, the event that the drawn ball is red. For Case 1, the probability of drawing a red ball,  $P_1(\text{Red})$ , is  $1/2$ ; for Case 2 the probability  $P_2(\text{Red})$  is  $1/40$ . Drawing a red ball is therefore considered *more informative* in Case 2 than in Case 1.

Another perspective of what makes the drawing of a red ball more informative in Case 2 is that  $P_2(\text{Red})=1/40$  indicates that the presence of a red ball in the Case 2 bag is a fact that will take a lot of trials on *average* to ascertain. On the other hand, it requires two trials, on average, to ascertain this fact in the Case 1 bag.

To summarize:

1. The information content of (or uncertainty associated with) the statement  $X = x_i$  increases as  $P(X = x_i)$  decreases;
2. The greater the dispersion of the distribution of possible values of  $X$ , the greater the uncertainty associated with the specific result that  $X = x_i$  and the lower the  $P(X = x_i)$ .

We now formalize this qualitative conceptual discussion.

### 10.1.2 Quantifying Information

For the discrete random variable  $X$ , let  $P(X = x_i) = f(x_i) = p_i$ ; define  $I(X = x_i)$  (or simply  $I(x_i)$ ) as the information content in the statement that the event  $X = x_i$  has occurred. From the discussion above, we know that  $I(x_i)$  should increase as  $p_i$  decreases, and vice versa. Formally, akin to the axioms of probability,  $I(x_i)$  must satisfy the following conditions:

1.  $I(x_i) \geq 0$ ; i.e. it must be non-negative;
2. For a *certain* event,  $I(x_i) = 0$ ;
3. For two stochastically independent random variables  $X, Y$ , let  $P(X = x_i) = p_i$  and  $P(Y = y_i) = q_i$  be the probabilities of the outcome of each indicated event; and let  $I(x_i)$  and  $I(y_i)$  be the respective information contents.

The total information content in the statement  $X = x_i$  and  $Y = y_i$  is the sum:

$$I(x_i, y_i) = I(x_i) + I(y_i) \quad (10.1)$$

This latter condition must hold because by their independence, the occurrence of one event has no effect on the occurrence of the other; as such one piece

of information is additional to the other, even though, also by independence, the probability of joint occurrence is the product:

$$P(X = x_i, Y = y_i) = p_i q_i \quad (10.2)$$

Claude Shannon in 1948 established that, up to a multiplicative constant, the desired unique measure of information content is defined by:<sup>1</sup>

$$I(X = x_i) = -\log_2 P(X = x_i) = -\log_2 f(x_i) \quad (10.3)$$

Note that this function satisfies all three conditions stated above.

## 10.2 Entropy

### 10.2.1 Discrete Random Variables

For the discrete random variable,  $X$ , and not just for a specific outcome  $X = x_i$ , Shannon suggests  $E[-\log_2 f(x)]$ , the “average” or mean information content, as a suitable measure of the information content in the pdf  $f(x)$ . This quantity, known as the *entropy* of the random variable, is defined by:

$$\mathcal{H}(X) = E[-\log_2 f(x)] = \sum_{i=1}^n [-\log_2 f(x_i)] f(x_i) \quad (10.4)$$

Expressed in this form,  $\mathcal{H}(X)$  has units of “bits” (for binary digits), a term that harks back to the original application for which the concepts were developed—a problem involving the characterization of the average minimum binary codeword length required to encode the output of an information source.

The expression for entropy is also sometimes written in terms of natural logarithms (with the matching, if whimsical, unit of “nats”) as:

$$\mathcal{H}(X) = E[-\ln f(x)] = \sum_{i=1}^n [-\ln f(x_i)] f(x_i) \quad (10.5)$$

One form differs from the other by only a multiplicative constant (specifically  $\ln 2$ ).

#### Example 10.1 ENTROPY OF A DETERMINISTIC VARIABLE

Compute the entropy of a variable  $X$  that takes on the value  $x_0$  with

<sup>1</sup>Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Tech. J.*, 27, 379-423 and 623-656.

probability 1, i.e. a deterministic variable that is always equal to  $x_0$ .

**Solution:**

Since for this variable,  $f(x_0) = 1$  and 0 otherwise, by definition,

$$\mathcal{H}(X) = -\log_2(1) \times 1 = 0 \quad (10.6)$$

so that the entropy of a deterministic variable is zero.

This example illustrates that when there is no uncertainty associated with a random variable, its entropy is zero.

**Example 10.2 ENTROPY OF A DISCRETE UNIFORM RANDOM VARIABLE**

Compute the entropy of the discrete uniform random variable,  $X \sim U_D(k)$ , whose pdf is given by:

$$f(x_i) = \frac{1}{k}; i = 1, 2, \dots, k. \quad (10.7)$$

**Solution:**

In this case

$$\mathcal{H}(X) = \sum_{i=1}^k \left[ -\log_2 \left( \frac{1}{k} \right) \right] \frac{1}{k} \quad (10.8)$$

$$= k \left[ -\log_2 \left( \frac{1}{k} \right) \right] \frac{1}{k} = \log_2 k \quad (10.9)$$

In this last example, note that:

1. In the limit as  $k \rightarrow \infty$  (i.e. the random variable can take any of an infinite number of possible discrete values),  $\mathcal{H}(X) \rightarrow \infty$ . Thus, as  $k \rightarrow \infty$  the uncertainty in  $X$  increases to the worst possible limit of complete uncertainty; the entropy also increases to match.
2. As  $k$  becomes smaller, uncertainty is reduced and  $\mathcal{H}(X)$  also becomes smaller; and when  $k = 1$ , the random variable becomes deterministic and  $\mathcal{H}(X) = 0$ , as obtained earlier in Example 10.1.

**Example 10.3 ENTROPY OF A BERNOULLI (BINARY) RANDOM VARIABLE**

Compute the entropy of the Bernoulli random variable whose pdf is given by:

$$f(x) = \begin{cases} 1-p & x = 0; \\ p & x = 1; \\ 0 & elsewhere. \end{cases} \quad (10.10)$$

**Solution:**

By definition, the entropy for this random variable is

$$\mathcal{H}(X) = -(1-p)\log_2(1-p) - p\log_2 p \quad (10.11)$$

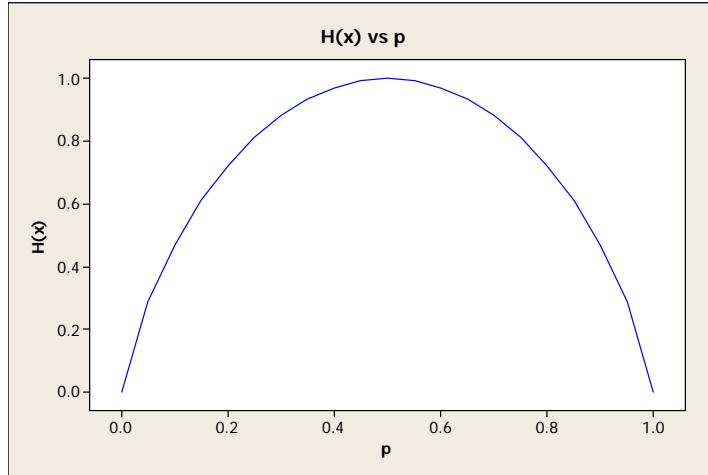


FIGURE 10.1: The entropy function of a Bernoulli random variable

This is a symmetric function of  $p$  which, as shown in the figure below, attains a maximum  $H^*(x) = 1$  when  $p = 0.5$ . Thus, the maximum entropy for a binary random variable is 1 bit, attained when the outcomes are equiprobable.

### 10.2.2 Continuous Random Variables

For continuous variables, the statement  $X = x_i$  is somewhat meaningless since it indicates an impossible event with a zero probability of occurrence. As defined in Eq (10.4) therefore, the entropy of *all* continuous random variables is infinite, which is not very useful. To extend the concept of entropy in a more useful manner to fundamentally continuous random variables thus requires additional considerations. Specifically, we must now introduce the concepts of quantization (or discretization) and “differential” entropy, as follows.

#### Quantization

For a continuous random variable  $X$  with a pdf  $f(x)$ , we know that if the set  $A = \{x : a < x < b\}$ , then  $P(A) = P(X \in A)$  is defined by:

$$P(A) = P(a < X < b) = \int_a^b f(x)dx \quad (10.12)$$

Let us now consider the case in which the interval  $[a, b]$  is divided into  $n$  subintervals of equal length  $\Delta x$ , so that the  $i^{th}$  subinterval  $A_i = \{x : x_i < x < x_i + \Delta x\}$ . Then, we recall that, for sufficiently small  $\Delta x$ , the probability

that  $X$  takes a value in this  $i^{th}$  subinterval is given by:

$$P(A_i) = P(x_i < X < (x_i + \Delta x)) \approx f(x_i)\Delta x \quad (10.13)$$

And since  $A$  is the union of  $n$  disjoint sets  $A_i$ , it follows from Eq (10.13) that:

$$P(A) = \sum_{i=1}^n P(A_i) \approx \sum_{i=1}^n f(x_i)\Delta x \quad (10.14)$$

from where we obtain the familiar result that in the limit as the quantization interval length  $\Delta x \rightarrow 0$ , the sum in (10.14) approaches the Riemann integral in (10.12), and in addition, the approximation error vanishes. But as it stands, the expression in (10.13) is a statement of the “differential” probability that  $X$  takes on a value in the differential interval between  $x_i$  and  $x_i + \Delta x$  for small, but non-zero,  $\Delta x$ .

Now, let  $Q(X)$  be a quantization function that places the continuous random variable anywhere in any one of the  $n$  quantized subintervals such that by  $Q = x_i$  we mean that  $x_i < X < x_i + \Delta x$ , or  $X \in A_i$ . Then  $P(Q = x_i)$  is given by:

$$P(Q = x_i) = P(x_i < X < x_i + \Delta x) \approx f(x_i)\Delta x \quad (10.15)$$

as in Eq (10.13). Since  $Q$  is discrete, we may compute its entropy as:

$$\mathcal{H}(Q) = \sum_{i=1}^n -\log_2 P(Q = x_i)P(Q = x_i) \quad (10.16)$$

so that, from Eq (10.15)

$$\begin{aligned} \mathcal{H}(Q) &\approx \sum_{i=1}^n -[\log_2 f(x_i)\Delta x]f(x_i)\Delta x \\ &= -\log_2(\Delta x) - \sum_{i=1}^n [\log_2 f(x_i)]f(x_i)\Delta x \end{aligned} \quad (10.17)$$

Some remarks are in order here:

1. In the limit as  $\Delta x \rightarrow 0$ , the sum in Eq (10.17) becomes (in many cases of practical interest) the integral

$$\tilde{\mathcal{H}}(X) = - \int_a^b [\log_2 f(x)]f(x)dx \quad (10.18)$$

but this is not enough to prevent the entropy  $\mathcal{H}(Q)$  from increasing without limit because of the  $\log_2(\Delta x)$  term. Thus, not surprisingly, the *exact* (not approximate) entropy of a continuous random variable again turns out to be infinite, as noted earlier.

2. For non-zero  $\Delta x$ , the entropy of the quantized version of the continuous random variable  $X$ , is finite, but there is residual quantization error; in the limit as  $\Delta x \rightarrow 0$ , the quantization error vanishes but then the entropy becomes infinite.
3. This implied trade-off between quantization accuracy and entropy of a continuous random variable is a well-known issue in information theory: *an infinite number of bits is required to specify a continuous random variable exactly; any finite-bit representation is achieved only at the expense of quantization error.*

### Differential Entropy

Let us now define a function  $h(X)$  such that

$$\mathcal{H}(X) = \log_2 h(X) \quad (10.19)$$

then it follows from Eq (10.17) that for the quantized random variable,

$$\log_2 h(Q) \approx -\log_2(\Delta x) - \sum_{i=1}^n [\log_2 f(x_i)] f(x_i) \Delta x \quad (10.20)$$

so that:

$$\log_2[h(Q)\Delta x] \approx \sum_{i=1}^n [\log_2 f(x_i)] f(x_i) \Delta x \quad (10.21)$$

In the limit as  $\Delta x \rightarrow 0$ , in the same spirit in which  $f(x)dx$  is considered the “differential” probability that  $X$  is in the interval  $[x, x+dx]$ , we may similarly define the “differential entropy”  $\tilde{\mathcal{H}}(X)$  of a continuous random variable,  $X$ , such that, by analogy with Eq (10.19),

$$\tilde{\mathcal{H}}(X) = \log_2[h(X)dx] \quad (10.22)$$

then we obtain the following result:

For the continuous random variable  $X$ , the *differential entropy* is defined as:

$$\tilde{\mathcal{H}}(X) = \int_{-\infty}^{\infty} [-\log_2 f(x)] f(x) dx \quad (10.23)$$

an expression that is reminiscent of the definition of the entropy of a discrete random variable given in (10.4), with the sum replaced by the integral. Thus, even though the entropy of a continuous random variable is infinite, the differential entropy is finite, and it is to entropy what  $f(x)dx$  is to probability. And just as the discrete pdf  $f(x_i)$  is fundamentally different from the continuous  $f(x)$ , so is the entropy function fundamentally different from the differential entropy function.

Nevertheless, even though the proper continuous variable counterpart of the discrete random variable’s entropy is the differential entropy defined in (10.23), we will still, for notational simplicity, use the same symbol  $\mathcal{H}(X)$  for

both, so long as we understand this to represent entropy in the discrete case and differential entropy when  $X$  is continuous. (We crave the indulgence of the reader to forgive this slight — but standard — abuse of notation.)

**Example 10.4 (DIFFERENTIAL) ENTROPY OF A CONTINUOUS UNIFORM RANDOM VARIABLE**

Compute the (differential) entropy of the continuous uniform random variable whose pdf is given by:

$$f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases} \quad (10.24)$$

**Solution:**

The differential entropy is given by

$$\mathcal{H}(X) = \int_a^b \left[ -\log_2 \frac{1}{c} \right] \frac{1}{c} dx \quad (10.25)$$

where  $c = (b - a)$ . This simplifies to

$$\mathcal{H}(X) = \log_2 c \quad (10.26)$$

which should be compared with the entropy obtained for the discrete uniform random variable in Example 10.2.

Some remarks about this example:

1. In the limit as  $a \rightarrow -\infty$  and  $b \rightarrow \infty$ , Eq (10.24) becomes a model for a random variable about which nothing is known, whose range of possible values is  $-\infty < x < \infty$ . For such a variable, even the differential entropy goes to infinity, since  $c \rightarrow \infty$  above in this case.
2. If we know that  $X$  is restricted to take values in a finite interval  $[a, b]$ , “adding” such constraining information reduces  $\mathcal{H}(X)$  to a finite value which depends on the interval length,  $c$ , as shown above. The longer this interval, (i.e. the more disperse  $f(x)$ ) the higher the (differential) entropy.
3. Thus in general, adding information about the random variable  $X$  reduces its entropy; conversely, to reduce entropy, one must add information.

We may now summarize the key points regarding the entropy of a (continuous or discrete) random variable  $X$ .

1. If nothing is known about a random variable,  $X$ , (making it infinitely uncertain) its entropy,  $\mathcal{H} = \infty$ ; if it is known perfectly with no uncertainty,  $\mathcal{H} = 0$ ;
2. Any information available about the behavior of the random variable in the form of some proper pdf,  $f(x)$ , reduces the entropy from the “absolute ignorance” state with  $\mathcal{H} = \infty$  to the finite, but still non-zero, entropy associated with the random variable’s pdf,  $f(x)$ .

We now discuss how these concepts can be used to obtain useful probability models in the face of incomplete knowledge.

### 10.3 Maximum Entropy Principles for Probability Modeling

When only limited information (perhaps in the form of its range, mean,  $\mu$ , variance,  $\sigma^2$ , or more generally, the expectation of some function  $G(X)$ ) is all that is available about a random variable  $X$ , clearly it is not possible to specify the full pdf  $f(x)$  uniquely because many pdf's exist that have the same range, or mean or variance, or whatever partial information has been supplied. The required full but unknown pdf contains additional information over and above what is legitimately known. To *postulate* a full pdf from such partial information therefore requires that we incorporate "extra" information to fill in what is missing. The problem at hand may be stated as follows:

How should we choose an appropriate  $f(x)$  to use in representing the random variable  $X$  given only a few of its characteristic parameters ( $\mu, \sigma^2$ , range, ...) and nothing else?

The "maximum entropy principle" states that the  $f(x)$  that adds the least amount of extra information, (i.e., the one with maximum entropy) should be chosen. The resulting pdf,  $f^*(x)$ , is then referred to as the "maximally unpresumptive distribution" because, of all the possible pdf's with the same characteristic parameters as those specified in the problem,  $f^*(x)$  is the least presumptive. Such a pdf is also called a "maximum entropy model" and the most common ones will now be derived.

### 10.4 Some Maximum Entropy Models

The procedure for obtaining maximum entropy models involves posing the optimization problem:

$$\max_{f(x)} \{ \mathcal{H}(X) = E[-\ln f(x)] \} \quad (10.27)$$

(where we have chosen the entropy function representation in “nats” for convenience) and solving it subject to the known information as constraints, as we now illustrate for several cases.

#### 10.4.1 Discrete Random Variable; Known Range

We begin with a random variable  $X$  for which the only known fact is that it can take on any of  $k$  discrete values  $x_1, x_2, \dots, x_k$ . What is an appropriate probability model for such a random variable?

Problem statement: Obtain  $f(x_i)$ ,  $i = 1, 2, \dots, k$ , given only that

$$\sum_{i=1}^k f(x_i) = 1. \quad (10.28)$$

The maximum entropy solution seeks to maximize

$$\mathcal{H}(X) = -\sum_{i=1}^k [\ln f(x_i)] f(x_i) \quad (10.29)$$

subject to Eq (10.28) as a constraint. This problem, and all subsequent ones, will be solved using principles of the calculus of variations (see for example<sup>2</sup>).

The Lagrangian functional for this problem is obtained as:

$$\Lambda(f) = \sum_{i=1}^k [-\ln f(x_i)] f(x_i) - \lambda \left[ \sum_{i=1}^k f(x_i) - 1 \right] \quad (10.30)$$

where  $\lambda$  is a Lagrange multiplier. The optimum  $f$  and the optimum value for  $\lambda$  are obtained from the Euler equations:

$$\frac{\partial \Lambda}{\partial f} = 0 \quad (10.31)$$

$$\frac{\partial \Lambda}{\partial \lambda} = 0 \quad (10.32)$$

where the second equation merely recovers the constraint. In the particular case at hand, we have

$$\frac{\partial \Lambda}{\partial f} = - \left[ \frac{1}{f(x_i)} f(x_i) + \ln f(x_i) \right] - \lambda = 0 \quad (10.33)$$

which yields, upon simplification,

$$f(x_i) = e^{-(1+\lambda)} = C \quad (10.34)$$

---

<sup>2</sup>Weinstock, R. (1974), *Calculus of Variations*, Dover Publications

$C$  is a constant to be determined such that Eq (10.28) is satisfied; i.e.

$$\sum_{i=1}^k f(x_i) = \sum_{i=1}^k C = 1 \quad (10.35)$$

so that

$$C = \frac{1}{k} \quad (10.36)$$

with the final result that the maximum entropy distribution for this discrete random variable  $X$  is:

$$f(x_i) = \frac{1}{k}; i = 1, 2, 3, \dots, k \quad (10.37)$$

Thus: *the maximum entropy principle assigns equal probabilities to each of the  $k$  outcomes of a discrete random variable when nothing else is known about the variable.*

This is a result in perfect keeping with intuitive common sense; in fact we have made use of it several times in our previous discussions. Note also that in Example 10.3, we saw that the Bernoulli random variable attains its maximum entropy for equiprobable outcomes.

#### 10.4.2 Discrete Random Variable; Known Mean

The random variable  $X$  in this case can take on discrete values  $1, 2, 3, \dots$ , and we seek an appropriate  $f(x_i)$  that is unknown except for

$$\sum_{i=1}^{\infty} f(x_i) = 1 \quad (10.38)$$

$$\sum_{i=1}^{\infty} x_i f(x_i) = \mu \quad (10.39)$$

by maximizing the entropy, subject to these two equations as constraints.

In this case, the Lagrangian is:

$$\Lambda(f) = \sum_{i=1}^{\infty} [-\ln f(x_i)] f(x_i) - \lambda_1 \left[ \sum_{i=1}^{\infty} f(x_i) - 1 \right] - \lambda_2 \left[ \sum_{i=1}^{\infty} x_i f(x_i) - \mu \right] \quad (10.40)$$

and the resulting Euler equations are:

$$\frac{\partial \Lambda}{\partial f} = -(\ln f(x_i) + 1) - \lambda_1 - \lambda_2 x_i = 0 \quad (10.41)$$

along with the other partial derivatives with respect to the Lagrange multipliers  $\lambda_1, \lambda_2$  that simply recover the two constraints. From Eq (10.41), we

obtain:

$$\begin{aligned} f(x_i) &= e^{-(1+\lambda_1)} (e^{-\lambda_2})^{x_i} \\ &= ab^{x_i} \end{aligned} \quad (10.42)$$

where the indicated constants  $a, b$  (functions of  $\lambda_1$  and  $\lambda_2$  as implied above) are determined by substituting (10.42) into the constraints equations, to obtain:

$$\sum_{x=1}^{\infty} ab^x = 1 \quad (10.43)$$

which, for  $|b| < 1$ , converges to give:

$$a \left( \frac{b}{1-b} \right) = 1 \quad (10.44)$$

Similarly,

$$\sum_{x=1}^{\infty} xab^x = \mu \quad (10.45)$$

yields:

$$a \frac{b}{(1-b)^2} = \mu \quad (10.46)$$

Solving (10.44) and (10.46) simultaneously for  $a$  and  $b$  now gives:

$$a = \frac{1}{\mu - 1} \quad (10.47)$$

$$b = \frac{\mu - 1}{\mu} \quad (10.48)$$

To tidy things up, if we now let  $p = 1/\mu$  (so that  $\mu = 1/p$ ), then

$$b = (1-p) \quad (10.49)$$

$$a = \frac{p}{1-p} \quad (10.50)$$

with the final result that the pdf we seek is given by:

$$f(x) = p(1-p)^{x-1} \quad (10.51)$$

which is immediately recognizable as the pdf of the geometric random variable. Thus: *the maximum entropy principle prescribes a geometric pdf for a discrete random variable with  $V_X = \{1, 2, 3, \dots, \infty\}$  and for which the mean  $\mu = 1/p$  is known.*

### 10.4.3 Continuous Random Variable; Known Range

In this case,  $X$  is a continuous random variable with  $V_X = \{x : a < x < b\}$  for which we seek an appropriate  $f(x)$  that is unknown except for:

$$\int_a^b f(x)dx = 1 \quad (10.52)$$

We will find  $f(x)$  by maximizing the (differential) entropy

$$\mathcal{H}(X) = \int_{-\infty}^{\infty} [-\ln f(x)]f(x)dx \quad (10.53)$$

subject to (10.52). The Lagrangian in this case takes the form:

$$\Lambda(f) = \int_a^b [-\ln f(x)]f(x)dx - \lambda \left( \int_a^b f(x)dx - 1 \right) \quad (10.54)$$

giving rise to the Euler equations:

$$\frac{\partial \Lambda}{\partial f} = -(\ln f(x) + 1) - \lambda = 0 \quad (10.55)$$

$$\frac{\partial \Lambda}{\partial \lambda} = 0 \quad (10.56)$$

again with the latter recovering the constraint. From (10.55) we obtain:

$$f(x) = e^{-(1+\lambda)} = c \quad (10.57)$$

a constant whose value is obtained from (10.52) as

$$\int_a^b cdx = c[b-a] = 1 \quad (10.58)$$

or,

$$c = \frac{1}{b-a} \quad (10.59)$$

Hence, the prescribed pdf is:

$$f(x) = \frac{1}{b-a}; a < x < b \quad (10.60)$$

which is recognizable as the continuous uniform pdf. This, of course, is the continuous version of the result obtained earlier for the discrete random variable encountered in Section 9.4.1.

#### 10.4.4 Continuous Random Variable; Known Mean

We seek a pdf  $f(x)$  that is unknown except for:

$$\int_{-\infty}^{\infty} f(x)dx = 1 \quad (10.61)$$

$$\int_{-\infty}^{\infty} xf(x)dx = \mu \quad (10.62)$$

The Lagrangian for maximizing the differential entropy in this case is:

$$\Lambda(f) = \int_{-\infty}^{\infty} [-\ln f(x)]f(x)dx - \lambda_1 \left( \int_{-\infty}^{\infty} f(x)dx - 1 \right) - \lambda_2 \left( \int_{-\infty}^{\infty} xf(x)dx - \mu \right) \quad (10.63)$$

The resulting Euler equations are:

$$\frac{\partial \Lambda}{\partial f} = -(\ln f(x) + 1) - \lambda_1 - \lambda_2 x = 0 \quad (10.64)$$

along with the constraints in (10.61) and (10.62). From (10.64) we obtain:

$$\begin{aligned} f(x) &= e^{-(1+\lambda_1)}e^{-\lambda_2 x} \\ &= C_1 e^{-\lambda_2 x} \end{aligned} \quad (10.65)$$

Substituting this back into (10.61) and (10.62) gives:

$$\begin{aligned} \int_{-\infty}^{\infty} C_1 e^{-\lambda_2 x} dx &= 1 \\ \int_{-\infty}^{\infty} C_1 x e^{-\lambda_2 x} dx &= \mu \end{aligned} \quad (10.66)$$

which, when solved simultaneously for  $C_1$  and  $\lambda_2$ , gives the result:

$$C_1 = 1/\mu; \lambda_2 = 1/\mu \quad (10.67)$$

for  $V_X = \{x : 0 \leq x < \infty\}$  (required for the integrals in (10.66) to be finite), so that (10.65) becomes:

$$f(x) = \begin{cases} \frac{1}{\mu} e^{-x/\mu} & x \geq 0 \\ 0 & otherwise \end{cases} \quad (10.68)$$

This is recognizable as the pdf of an exponential random variable (the continuous version of the result obtained earlier in section 9.4.2). Thus: *the maximum entropy principle prescribes an exponential pdf for the continuous random variable for which only the mean is known.*

#### 10.4.5 Continuous Random Variable; Known Mean and Variance

We seek a pdf  $f(x)$  that is unknown except for:

$$\int_{-\infty}^{\infty} f(x)dx = 1 \quad (10.69)$$

$$\int_{-\infty}^{\infty} xf(x)dx = \mu \quad (10.70)$$

$$\int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx = \sigma^2 \quad (10.71)$$

Once more, we work with the Lagrangian, in this case:

$$\begin{aligned} \Lambda(f) &= \int_{-\infty}^{\infty} [-\ln f(x)]f(x)dx - \lambda_1 \left( \int_{-\infty}^{\infty} f(x)dx - 1 \right) - \lambda_2 \left( \int_{-\infty}^{\infty} xf(x)dx - \mu \right) \\ &\quad - \lambda_3 \left( \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx - \sigma^2 \right) \end{aligned} \quad (10.72)$$

The Euler equations are:

$$\frac{\partial \Lambda}{\partial f} = -\ln f(x) - 1 - \lambda_1 - \lambda_2 x - \lambda_3(x - \mu)^2 = 0 \quad (10.73)$$

along with the three constraints in (10.69 – 10.71). Solving (10.73) gives:

$$f(x) = C_1 e^{-\lambda_2 x} e^{-\lambda_3(x - \mu)^2} \quad (10.74)$$

Substituting this back into the constraints and using the result:

$$\int_{-\infty}^{\infty} e^{-au^2} = \sqrt{\frac{\pi}{a}} \quad (10.75)$$

upon some algebraic manipulations, yields:

$$C_1 = \frac{1}{\sigma\sqrt{2\pi}} \quad (10.76)$$

$$\lambda_2 = 0; \quad (10.77)$$

$$\lambda_3 = \frac{1}{2\sigma^2} \quad (10.78)$$

giving as the final result:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left[ \frac{-(x-\mu)^2}{2\sigma^2} \right]} \quad (10.79)$$

the familiar Gaussian pdf. Thus, *when only the mean,  $\mu$ , and the variance  $\sigma^2$  are all that we legitimately know about a continuous random variable, the maximally unpresumptive distribution  $f(x)$  is the Gaussian pdf.*

#### 10.4.6 Continuous Random Variable; Known Range, Mean and Variance

We present, without proof or detailed derivations, that for  $X$  continuous, given the range  $(0,1)$ , the mean,  $\mu$ , and variance,  $\sigma^2$ , the maximum entropy model is the beta  $B(\alpha, \beta)$  pdf:

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad (10.80)$$

### 10.5 Maximum Entropy Models from General Expectations

#### 10.5.1 Single Expectations

Suppose that  $X$  is a random variable (discrete or continuous) defined on the space  $V_X$  and whose pdf  $f(x)$  is unknown, except that  $E[G(X)]$ , the expected value of some (continuous) function  $G(X)$ , is some known constant,  $\gamma$ , i.e.

$$\sum_{i=1}^n G(x_i)f(x_i) = \gamma \quad (10.81)$$

for discrete  $X$ , or

$$\int_{-\infty}^{\infty} G(x)f(x)dx = \gamma \quad (10.82)$$

for continuous  $X$ . Given this single piece of information, expressions for the resulting maximum entropy models for any  $G(X)$  will now be derived.

#### Discrete Case

The Lagrangian in the discrete case is given by:

$$\Lambda(f) = \sum_{i=1}^n [-\ln f(x_i)]f(x_i) - \lambda \left[ \sum_{i=1}^n f(x_i)G(x_i) - \gamma \right] \quad (10.83)$$

which is easily rearranged to give:

$$\Lambda(f) = - \sum_{i=1}^n f(x_i) \ln \left[ \frac{f(x_i)}{Ce^{-\lambda G(x_i)}} \right] \quad (10.84)$$

It can be shown that as presented in Eq (10.84),  $\Lambda(f) \leq 0$ , attaining its maximum value of 0 when

$$f(x_i) = Ce^{-\lambda G(x_i)} \quad (10.85)$$

This, then, is the desired maximum entropy model for any function  $G(X)$  of the discrete random variable,  $X$ ; the constants  $C$  and  $\lambda$  are determined such that  $f(x_i)$  satisfies the constraint:

$$\sum_{i=1}^{\infty} f(x_i) = 1 \quad (10.86)$$

and the supplied expectation information in Eq (10.81).

### Continuous Case

Analogous to the discrete case, the Lagrangian in the continuous case is:

$$\Lambda(f) = \int_{-\infty}^{\infty} [-\ln f(x)]f(x)dx - \lambda \left( \int_{-\infty}^{\infty} f(x)G(x)dx - \gamma \right) \quad (10.87)$$

which again rearranges to:

$$\Lambda(f) = - \int_{-\infty}^{\infty} f(x) \ln \left[ \frac{f(x)}{Ce^{-\lambda G(x)}} \right] dx \quad (10.88)$$

It can also be shown that  $\Lambda(f)$  in Eq (10.88) is maximized when:

$$f(x) = Ce^{-\lambda G(x)} \quad (10.89)$$

Again, this represents the maximum entropy model for any function  $G(X)$  of the continuous random variable  $X$ , with the indicated constants to be determined to satisfy the constraint:

$$\int_{-\infty}^{\infty} f(x)dx = 1 \quad (10.90)$$

and the given expectation in Eq (10.82).

It is a simple enough exercise (see Exercises 10.8 and 10.9) to establish the following results and hence confirm, from this alternative route, the maximum entropy results presented earlier in section 10.4:

1. If  $G(X) = 0$ , indicating that *nothing* is known about the random variable, the pdf prescribed by Eq (10.85) is the discrete uniform distribution, and by Eq (10.89), the continuous uniform distribution.
2. If  $G(X) = X$ , so that the supplied information is the mean value, for discrete  $X$ , the pdf prescribed by Eq (10.85) is the geometric distribution; for continuous  $X$ , the pdf prescribed by Eq (10.89) is the exponential distribution.
3. If  $X$  is continuous and

$$G(X) = (X - \mu)^2 \quad (10.91)$$

then the pdf prescribed by Eq (10.89) is the Gaussian distribution.

### 10.5.2 Multiple Expectations

In the event that information is available about the expectations of  $m$  functions  $G_j(X); j = 1, 2, \dots, m$ , of the random variable, i.e.

$$\sum_{i=1}^n G_j(x_i) f(x_i) = \gamma_j; j = 1, 2, \dots, m; \quad (10.92)$$

for discrete  $X$ , or

$$\int_{-\infty}^{\infty} G_j(x) f(x) dx = \gamma_j; j = 1, 2, \dots, m; \quad (10.93)$$

for continuous  $X$ , so that in each case,  $\gamma_j$  are known constants, then the Lagrangians are obtained as

$$\Lambda(f) = \sum_{i=1}^n [-\ln f(x_i)] f(x_i) - \sum_{j=1}^m \lambda_j \left[ \sum_{i=1}^n f(x_i) G_j(x_i) - \gamma_j \right] \quad (10.94)$$

$$\Lambda(f) = \int_{-\infty}^{\infty} [-\ln f(x)] f(x) dx - \sum_{j=1}^m \lambda_j \left[ \int_{-\infty}^{\infty} f(x) G_j(x) dx - \gamma_j \right] \quad (10.95)$$

It can be shown that these Lagrangians are maximized for pdf's given by:

$$f(x_i) = C e^{-[\sum_{j=1}^m \lambda_j G_j(x_i)]} \quad (10.96)$$

for discrete  $X$ , and

$$f(x) = C e^{-[\sum_{j=1}^m \lambda_j G_j(x)]} \quad (10.97)$$

for continuous  $X$ , generalizing the results in Eqs (10.85) and (10.89). These results are from a theorem by Boltzmann (1844–1906), the Austrian theoretical physicists credited with inventing statistical thermodynamics and statistical mechanics. The constant  $C$  in each case is the normalizing constant determined such that  $\int f(x) dx$  and  $\sum f(x_i)$  equal 1; the  $m$  Lagrange multipliers  $\lambda_1, \lambda_2, \dots, \lambda_m$  are obtained from solving simultaneously the  $m$  equations representing the known expectations in Eqs (10.92) and (10.93).

The following are two applications of this set of results. Consider the case where, for a continuous random variable  $X$ ,

$$G_1(X) = X \quad (10.98)$$

$$G_2(X) = \ln X \quad (10.99)$$

and

$$E[G_1(X)] = \alpha; \alpha > 0 \quad (10.100)$$

$$E[G_2(X)] = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \quad (10.101)$$

then the pdf prescribed by Eq (10.97) is:

$$f(x) = Ce^{[-\lambda_1 x - \lambda_2 \ln x]} = Cx^{-\lambda_2} e^{-\lambda_1 x} \quad (10.102)$$

and upon evaluating the constants, we obtain:

$$f(x) = \frac{1}{\Gamma(\alpha)} e^{-x} x^{\alpha-1} \quad (10.103)$$

recognizable as the pdf for the Gamma random variable.

For another continuous random variable  $X$ , with

$$G_1(X) = \ln X \quad (10.104)$$

$$G_2(X) = \ln(1 - X) \quad (10.105)$$

and

$$E[G_1(X)] = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - \frac{\Gamma'(\alpha + \beta)}{\Gamma(\alpha + \beta)} \quad (10.106)$$

$$E[G_2(X)] = \frac{\Gamma'(\beta)}{\Gamma(\beta)} - \frac{\Gamma'(\alpha + \beta)}{\Gamma(\alpha + \beta)} \quad (10.107)$$

then the pdf prescribed in Eq (10.97) is:

$$f(x) = Ce^{[-\lambda_1 \ln x - \lambda_2 \ln(1-x)]} = Cx^{-\lambda_1}(1-x)^{-\lambda_2} \quad (10.108)$$

and upon evaluating the constants, we obtain:

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad (10.109)$$

again recognizable as the pdf for a Beta random variable.

## 10.6 Summary and Conclusions

This chapter has been concerned with the problem of how to determine appropriate (and complete) probability models when only partial information is available about the random variable in question. The first-principles approach to probability model development discussed in the earlier chapters (Chapter 8 for discrete random variables and Chapter 9 for the continuous type), is predicated on the availability of complete phenomenological information about the random variable of interest. When this is not the case, and only partial information is available, model development must be approached differently. This chapter has offered such an alternative approach—one based on the “maximum entropy principle.” The essence of this principle is that of all

the several pdfs whose characteristics are consistent with the available partial information, the one that adds the least amount of extraneous information—the least presumptive—should be chosen as an appropriate model. To realize this intuitively appealing concept *fully* in practice, of course, requires advanced optimization theory, much of which is not expected to be familiar to the average reader. Still, enough of the derivation details have been presented to allow the reader to appreciate how the results came to be.

It is interesting to note now in retrospect that *all* the results presented in this chapter have involved familiar pdfs encountered previously in earlier discussions. This should not give the impression that these are the only useful maximum entropy distributions; neither should this be construed as implying that all pdfs encountered previously have maximum entropy interpretations. The scope of coverage was designed first to demonstrate to (and inspire confidence in) the reader that this approach, even though somewhat esoteric, does in fact lead to results that “make sense.” Secondly, this coverage was also designed to offer a different perspective of some of these familiar models. For instance, as a model for residence time distribution in chemical reaction engineering, we have seen the exponential distribution arise from chemical engineering arguments (Chapter 2), probability arguments (Chapter 9), and now from maximum entropy considerations. The same is true for the geometric distribution as a model for polymer chain length distribution (see Application Problem 10.12). But the application of this principle in practice extends well beyond the catalog of familiar results shown here, for example, see Phillips *et al.*, (2004)<sup>3</sup> for an application to the problem of modeling geographic distributions of species, a critical problem in conservation biology.

With the discussion in this chapter behind us, we have now completed our study of probability models and their development. The discussion in the next chapter is a case study illustrating how probability models are developed, validated and applied in solving the complex and important practical problem of optimizing the effectiveness of in-vitro fertilization.

The main points and results of this chapter are summarized in Table 10.1.

## REVIEW QUESTIONS

1. What are the three axioms employed in quantifying the information content of the statement  $P(X = x_i) = p_i$ ?
2. In what ways are the axioms of “information content” akin to the axioms of probability encountered in Chapter 4?
3. What is the entropy of a discrete random variable  $X$  with a pdf  $f(x)$ ?

---

<sup>3</sup>S. J. Phillips, M. Dudík and R. E. Schapire, (2004) “A Maximum Entropy Approach to Species Distribution Modeling,” *Proc. Twenty-First International Conference on Machine Learning*, 655-662.

**TABLE 10.1:** Summary of maximum entropy probability models

| Known<br>Random Variable<br>Characteristics   | Maximum<br>Entropy<br>Distribution                         | Probability<br>Model   |
|---|--|--|
| Discrete<br>Binary (0,1)  | Bernoulli<br>$Bn(0.5)$                                     | $f(0) = 0.5$<br>$f(1) = 0.5$   |
| Discrete<br>Range: $i = 1, 2, \dots, k$   | Uniform<br>$U_D(k)$  | $f(x_i) = \frac{1}{k}$   |
| Discrete<br>Mean, $\mu$   | Geometric<br>$G(p); p = \frac{1}{\mu}$                     | $f(x) = p(1-p)^{x-1}$  |
| Continuous<br>Range: $a \leq x \leq b$  | Uniform<br>$U(a, b)$                                       | $f(x) = \frac{1}{(b-a)}$   |
| Continuous<br>Mean, $\mu$   | Exponential<br>$\mathcal{E}(\beta); \beta = \frac{1}{\mu}$ | $f(x) = \frac{1}{\beta} e^{-x/\beta}$  |
| Continuous<br>Mean, $\mu$ ; Variance, $\sigma^2$  | Gaussian<br>$N(\mu, \sigma^2);$                            | $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$           |
| Continuous<br>Mean, $\mu$ ; Variance, $\sigma^2$<br>Range: $0 \leq x \leq 1$  | Beta<br>$B(\alpha, \beta);$                                | $f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$ |
| $G(X) = 0; a \leq x \leq b$<br>$E[G(X)] = 0$  | Uniform<br>$U(a, b)$                                       | $f(x) = \frac{1}{(b-a)}$   |
| $G_1(X) = X; G_2(X) = \ln X$<br>$E[G_1(X)] = \alpha$<br>$E[G_2(X)] = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$  | Gamma<br>$\gamma(\alpha, 1)$                               | $f(x) = \frac{1}{\Gamma(\alpha)} e^{-x} x^{\alpha-1}$  |
| $G_1(X) = \ln X$<br>$G_2(X) = \ln(1-X)$<br>$E[G_1(X)] = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - \frac{\Gamma'(\alpha+\beta)}{\Gamma(\alpha+\beta)}$<br>$E[G_2(X)] = \frac{\Gamma'(\beta)}{\Gamma(\beta)} - \frac{\Gamma'(\alpha+\beta)}{\Gamma(\alpha+\beta)}$ | Beta<br>$B(\alpha, \beta);$                                | $f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$ |

4. Why is entropy as defined for discrete random variables not very useful for continuous random variables?
5. What is the corresponding entropy concept for continuous random variables?
6. What is quantization and why must there always be a trade-off between quantization accuracy and entropy of a continuous random variable?
7. What is the differential entropy of a continuous random variable?
8. What is the entropy of a random variable about which nothing is known?
9. What is the entropy of a variable with no uncertainty?
10. What effect does any additional information about a random variable have on its entropy?
11. Provide a succinct statement of the primary problem of this chapter.
12. What is the “maximum entropy principle” for determining full pdfs when only partial information is available?
13. What is the maximum entropy distribution for a discrete random variable  $X$  for which the only known fact is that it can take on any of  $k$  discrete values  $x_1, x_2, \dots, x_k$ ?
14. What is the maximum entropy distribution for a discrete random variable  $X$  for which the mean is known and nothing else?
15. What is the maximum entropy distribution for a continuous random variable  $X$  for which the only known fact is its range,  $V_X = \{x : a < x < b\}$ ?
16. What is the maximum entropy distribution for a continuous random variable  $X$  for which the mean is known and nothing else?
17. What is the maximum entropy distribution for a continuous random variable  $X$  for which the mean,  $\mu$ , and variance,  $\sigma^2$ , are known and nothing else?
18. What is the maximum entropy distribution for a continuous random variable  $X$  for which the range,  $(0,1)$ , mean,  $\mu$ , and variance,  $\sigma^2$ , are known and nothing else?
19. Which two equations arise from a theorem of Boltzmann and how are they used to obtain maximum entropy distributions?

## EXERCISES

**10.1** Using the principles of differential calculus, establish that the entropy of the Bernoulli random variable, shown in Eq (10.11), i.e.,

$$\mathcal{H}(X) = -(1-p)\log_2(1-p) - p\log_2 p$$

is maximized when  $p = 0.5$ .

**10.2** Determine the maximum entropy distribution for the Binomial random variable  $X$ , the total number of successes in  $n$  Bernoulli trials, when all that is known is that with each trial, there are exactly only two outcomes. (Hint:  $X = \sum_{i=1}^n X_i$ , where  $X_i$  is a Bernoulli random variable.)

**10.3** Determine the entropy for the geometric random variable,  $G(p)$ , with the pdf

$$f(x) = pq^{x-1}; x = 1, 2, \dots$$

and compare it to the entropy obtained for the Bernoulli random variable in Example 10.3 in the text.

**10.4** Show that the entropy for the exponential random variable,  $\mathcal{E}(\beta)$ , with pdf

$$f(x) = \frac{1}{\beta}e^{-x/\beta}; 0 < x < \infty$$

is given by:

$$\mathcal{H}(X) = 1 + \ln \beta \quad (10.110)$$

**10.5** Show that the entropy for the Gamma random variable,  $\gamma(\alpha, \beta)$ , with pdf

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} e^{-x/\beta} x^{\alpha-1}; 0 < x < \infty$$

is given by:

$$\mathcal{H}(X) = \alpha + \ln \beta + \ln \Gamma(\alpha) + (1-\alpha) \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \quad (10.111)$$

Directly from this result, write an expression for the entropy of the  $\chi^2(r)$  random variable.

**10.6** Show that the entropy for the Gaussian  $N(\mu, \sigma^2)$  random variable with pdf

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

is given by:

$$\mathcal{H}(X) = \ln (\sigma \sqrt{2\pi e}) \quad (10.112)$$

and hence establish that the entropy of a Gaussian random variable depends only on  $\sigma$  and not  $\mu$ . Why does this observation make sense? In the limit as  $\sigma \rightarrow \infty$ , what happens to the entropy,  $\mathcal{H}(X)$ ?

**10.7** Show that the entropy for the Lognormal  $\mathcal{L}(\alpha, \beta^2)$  random variable with pdf

$$f(x) = \frac{1}{x\beta\sqrt{2\pi}} \exp\left\{-\frac{(\ln x - \alpha)^2}{2\beta^2}\right\}; 0 < x < \infty$$

is given by:

$$\mathcal{H}(X) = \ln(\beta\sqrt{2\pi e}) + \alpha \quad (10.113)$$

Compare this with the expression for the entropy of the Gaussian random variable in Exercise 9.4, Eq (10.112). Why does the entropy of the Lognormal random variable depend linearly on  $\alpha$  while the entropy of the Gaussian random variable does not depend on the corresponding parameter,  $\mu$  at all?

**10.8** The maximum entropy distribution for a random variable  $X$  for which  $G(X)$  and its expectation,  $E[G(X)]$ , are specified, was given in Eq (10.85) for discrete  $X$ , and Eq (10.89) for continuous  $X$  i.e.,

$$f(x) = \begin{cases} Ce^{-\lambda G(x_i)}; & \text{for discrete } X \\ Ce^{-\lambda G(x)}; & \text{for continuous } X \end{cases}$$

Determine  $f(x)$  completely (i.e., determine  $C$  and  $\lambda$  explicitly) under the following conditions:

- (i)  $G(X_i) = 0$ ;  $i = 1, 2, \dots, k$ , a discrete random variable for which nothing is known except its range;
- (ii)  $G(X) = 0$ ;  $a < X < b$ ;
- (iii)  $G(X_i) = X_i$ ;  $i = 1, 2, \dots$ ;  $E[G(X)] = \mu$
- (iv)  $G(X) = X$ ;  $0 < x < \infty$ ;  $E[G(X)] = \mu$

**10.9** For the continuous random variable  $X$  for which

$$G(X) = (X - \mu)^2$$

is specified along with its expectation,

$$E[G(X)] = E[(X - \mu)^2] = \sigma^2,$$

the maximum entropy distribution was given in Eq (10.89) as:

$$f(x) = Ce^{-\lambda G(x)};$$

Show that the constants in this pdf are given by:

$$C = \frac{1}{\sigma\sqrt{2\pi}} \quad (10.114)$$

$$\lambda = \frac{1}{2\sigma^2} \quad (10.115)$$

and thus establish the result that:

*the maximally unpresumptive distribution for a random variable  $X$  for which only the mean,  $\mu$ , and variance,  $\sigma^2$ , are legitimately known, is the Gaussian pdf:*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right] \quad (10.116)$$

You may find the following identity useful:

$$\int_{-\infty}^{\infty} e^{-au^2} du = \sqrt{\frac{\pi}{a}} \quad (10.117)$$

**10.10** Given the following information about a continuous random variable,  $X$ ,

$$G_1(X) = X; \text{ and } G_2(X) = \ln X$$

along with

$$E[G_1(X)] = \alpha; \alpha > 0; \text{ and } E[G_2(X)] = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$$

it was stated in the text that the maximum entropy pdf prescribed by Eq (10.97) is:

$$f(x) = Ce^{[-\lambda_1 x - \lambda_2 \ln x]} = Cx^{-\lambda_2} e^{-\lambda_1 x} \quad (10.118)$$

Determine the constants  $C, \lambda_1$  and  $\lambda_2$  and hence establish the result given in Eq (10.103).

**10.11** Revisit Exercise 10.10 for the case where the information available about the random variable  $X$  is:

$$G_1(X) = \frac{X}{\beta}; \text{ and } G_2(X) = \ln\left(\frac{X}{\beta}\right)$$

along with

$$E[G_1(X)] = \alpha; \alpha > 0; \text{ and } E[G_2(X)] = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$$

obtain an explicit expression for the maximum entropy pdf in this case.

## APPLICATION PROBLEMS

**10.12** In certain polymerization processes, the polymer product is made by the sequential addition of monomer units to a growing chain. At each step, after the chain has been initiated, a new monomer may be added, propagating the chain, or a termination event can occur, stopping the growth; whether the growing chain propagates or terminates is random. The random nature of the propagation and termination events is responsible for polymer products having chains of variable lengths. As such, because it is a count of the number of monomer units in the chain,  $X$ , the length of a particular polymer chain, is a discrete random variable.

Now consider the case where the only information available about a particular process is the kinetic rate of the termination reaction, given as  $R_T$  per min, which can be interpreted as implying that an average of  $R_T$  chain terminations occur per min. By considering the reciprocal of  $R_T$ , i.e.,

$$p = \frac{1}{R_T}$$

as the probability that a termination reaction will occur, obtain  $f(x)$ , a maximum

entropy distribution for the polymer chain length, in terms of  $p$ . This distribution is often known as the “most probable” chain length distribution.

**10.13** As introduced very briefly in Chapter 2, the continuous stirred tank reactor (CSTR), a ubiquitous equipment used in the chemical industry to carry out a wide variety of chemical reactions, consists of a tank of volume  $V$  liters, through which the reactant stream flows continuously at a rate of  $F$  liters/sec; the content is vigorously stirred to ensure uniform mixing, and the product is continuously withdrawn from the outlet at the same rate,  $F$  liters/sec. Because of the vigorous mixing, the amount of time any particular fluid element spends in the reactor—the reactor residence time—varies randomly, so that there is in fact not a single value for residence time,  $X$ , but a distribution of values. Clearly, the residence time affects the productivity of the reactor, and characterizing it is a central concern in chemical reaction engineering.

Now, a stream continuously fed at a rate  $F$  liters/sec through a reactor of volume  $V$  liters implies an average residence time,  $\tau$  in secs, given by

$$\tau = \frac{V}{F}$$

Given only this information, obtain a maximum entropy distribution for the residence time in the CSTR, and compare it with the result in Section 2.1.2 of Chapter 2.

**10.14** Integrins are transmembrane receptors that link the actin cytoskeleton of a cell to the extra cellular matrix (ECM). This connection, which constantly and dynamically reorganizes in response to mechanical, chemical, and other environmental cues around the cell, leads to lateral assembly of integrins into small stationary “focal complexes” or clusters of integrins. Integrin clustering, an extremely important process in cell attachment and migration, is a stochastic process that results in heterogeneous populations of clusters that are best characterized with distributions. One of the many characteristics of an integrin cluster is its shape. Because integrin clusters grow or shrink in different directions depending on the orientation and tension of the actin cytoskeleton, the shape of an integrin cluster provides useful information concerning the forces acting on a particular adhesive structure.

The shape of integrin clusters is often idealized as an ellipse and quantified by its eccentricity,  $\varepsilon$ , the ratio of the distance between the foci of the representative ellipse to the length of its major axis. This quantity has the following properties:

1. It is scaled between 0 and 1, i.e.,  $0 \leq \varepsilon \leq 1$ ;
2.  $\varepsilon \approx 1$  for elongated clusters; for circular clusters,  $\varepsilon \approx 0$ ;
3. Physiologically, integrin clusters in *adherent cells* tend to be more elongated than circular; non-adherent cells tend to have more circular integrin clusters.

Given that the average and variance of cluster eccentricity is often known for a particular cell (and, in any event, these can be determined experimentally), obtain a maximum entropy distribution to use in representing this aspect of integrin clustering.

Data obtained by Welf (2009)<sup>4</sup> from Chinese Hamster Ovary (CHO) cells stably expressing the integrin  $\alpha IIb\beta 3$ , indicated an average eccentricity  $\mu_\varepsilon = 0.92$  and

---

<sup>4</sup>Welf, E. S. (2009). *Integrative Modeling of Cell Adhesion Processes*, PhD Dissertation, University of Delaware.

variance,  $\sigma_\varepsilon^2 = 0.003$  for the particular collection of integrin clusters studied. From this information, obtain a specific theoretical pdf that characterizes the size distribution of this experimental population of integrin clusters and determine the mode of the distribution. Plot the pdf; and from the shape of this distribution, comment on whether you expect the specific clusters under study to belong to adherent or non-adherent cells.

**10.15** Mee (1990)<sup>5</sup> presented the following data on the wall thickness (in ins) of cast aluminum cylinder heads used in aircraft engine cooling jackets. The mean and variance of the wall thickness are therefore considered as known. If a full pdf is to be prescribed to characterize this important property of the manufactured cylinder heads, use the maximum entropy principle to postulate one. Even though there are only 18 data points, plot the theoretical pdf versus a histogram of the data and comment on the model fit.

|       |       |       |       |       |       |       |       |       |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 0.223 | 0.228 | 0.214 | 0.193 | 0.223 | 0.213 | 0.218 | 0.215 | 0.233 |
| 0.201 | 0.223 | 0.224 | 0.231 | 0.237 | 0.217 | 0.204 | 0.226 | 0.219 |

**10.16** The total number of occurrences of a rare event in an interval of time  $(0, T)$ , when the event occurs at a mean rate,  $\eta$  per unit time, is known to be a Poisson random variable. However, given that an event has occurred in this interval, and without knowing exactly when the event occurred, determine a maximum entropy distribution for the time of occurrence of this lone event within this interval.

Now let  $X$  be the time of occurrence in the normalized unit time interval,  $(0,1)$ . Using the just-obtained maximum entropy distribution, derive the distribution for the log-transformed variable,

$$Y = -\frac{1}{\eta} \ln X \quad (10.119)$$

Interpret your result in terms of what you know about the Poisson random variable and the inter-arrival times of Poisson events.

---

<sup>5</sup>Mee, R. W., (1990). "An improved procedure for screening based on a correlated, normally distributed variable," *Technometrics*, 32, 331–337.

# Chapter 11

## ***Application Case Studies II: In-Vitro Fertilization***

|   |   |     |
|---|---|-----|
| 11.1  | Introduction .....  | 364 |
| 11.2  | In-Vitro Fertilization and Multiple Births .....                            | 365 |
| 11.2.1  | Background and Problem Definition .....                                     | 365 |
| 11.2.2  | Clinical Studies and Recommended Guidelines .....                           | 367 |
| Factors affecting live-birth and multiple-birth rates ..... | 367   |     |
| Prescriptive Studies .....                                  | 368   |     |
| Determining Implantation Potential .....                    | 368   |     |
| Qualitative Optimization Studies .....                      | 369   |     |
| 11.3  | Probability Modeling and Analysis .....                                     | 371 |
| 11.3.1  | Model Postulate .....   | 371 |
| 11.3.2  | Prediction .....  | 372 |
| 11.3.3  | Estimation .....  | 373 |
| 11.4  | Binomial Model Validation .....   | 375 |
| 11.4.1  | Overview and Study Characteristics .....                                    | 375 |
| 11.4.2  | Binomial Model versus Clinical Data .....                                   | 377 |
| 11.5  | Problem Solution: Model-based IVF Optimization and Analysis .....           | 384 |
| 11.5.1  | Optimization .....  | 385 |
| 11.5.2  | Model-based Analysis .....  | 386 |
| 11.5.3  | Patient Categorization and Theoretical Analysis of Treatment Outcomes ..... | 387 |
| 11.6  | Sensitivity Analysis .....  | 392 |
| 11.6.1  | General Discussion .....  | 392 |
| 11.6.2  | Theoretical Sensitivity Analysis .....                                      | 394 |
| 11.7  | Summary and Conclusions .....   | 395 |
| 11.7.1  | Final Wrap-up .....   | 395 |
| 11.7.2  | Conclusions and Perspectives on Previous Studies and Guidelines .....       | 397 |
|   | References .....  | 398 |
|   | PROJECT ASSIGNMENT .....  | 399 |

*It is a test of true theories  
not only to account for but to predict phenomena*

William Whewell (1794–1866)

“The mathematician,” Tobias Dantzig (1884–1956) noted with pitch-perfect precision in his captivating book, *Number*, “may be compared to a designer of garments, who is utterly oblivious of the creature whom his garments may fit. To be sure, his art originated in the necessity for clothing such creatures, but this was long ago; to this day a shape will occasionally appear which will fit into the garment as if the garment had been made for it. Then there is no end of surprise and of delight!”

Such a shape appears in this chapter in the form of *in-vitro fertilization*

(IVF), an iconic 20<sup>th</sup> century “creature” of whose future existence the designers of the garments of probability distributions (specifically, the binomial distribution) were utterly oblivious a few centuries ago. Our surprise and delight do not end upon discovering just how well the binomial pdf model fits, as if custom-made for IVF analysis; there is the additional and completely unexpected bonus discovery that, with no modification or additional embellishments, this model also is perfectly suited to solving the vexing problem of maximizing the chances of success while simultaneously minimizing the risk of multiple births *and* total failure. This chapter is the second in the series of case studies designed to illustrate how the probabilistic framework can be used effectively to solve complex, real-life problems involving randomly varying phenomena.

---

## 11.1 Introduction

*“When Theresa Anderson learned that she was pregnant with quintuplets, she was dumbfounded. She had signed up to be a surrogate mother for an infertile couple and during an in vitro fertilization procedure doctors introduced five embryos into her womb. ‘They told me there was a one in 30 chance that one would take,’ she says. Instead, all five took. The 26-year-old mother of two endured a difficult pregnancy and delivered the boys in April [2005] at a Phoenix hospital. The multiple births made headlines across the country as a feel-good tale. But they also underscore a reality of the fertility business: Many clinics are implanting more than the recommended number of embryos in their patients, raising the risks for women.”*

So began an article by Sylvia Pagán Westphal that appeared in the Wall Street Journal (WSJ) on October 7, 2005. In-vitro fertilization (IVF), the very first of a class of procedures collectively known as “Assisted Reproductive Technology” (ART), was originally developed specifically to treat infertility caused by blocked or damaged fallopian tubes; it is now used to treat a variety of infertility problems, with impressive success. With IVF, eggs and sperm are combined in a laboratory to fertilize “in vitro” (literally “in glass”). The fertilized eggs are later transferred into the woman’s uterus, where, in the successful cases, implantation, embryo development, and ultimately, a live birth, will occur as with all other normal pregnancies. Since 1978 when the first so-called “test-tube baby” was born, IVF has enabled an increasing number of otherwise infertile couples to experience the joy of having children. So successful in fact has IVF been that its success rates now compare favorably to natural pregnancy rates in any given month, especially when the woman is under 40 years of age and there are no sperm problems.

With the advent of *oocyte donation* (Yaron, et al, 1997; Reynolds, et al, 2001), where the eggs used for IVF have been donated typically by younger

women, the once formidable barriers to success due to age or ovarian status are no longer as serious. Today, ART with oocyte donation is one of the most successful treatment programs for infertility. Recent studies have reported pregnancy rates as high as 48% per retrieval: out of a total of 6,936 IVF procedures using donated oocytes carried out in 1996 and 1997, 3,320 resulted in pregnancies and 2,761 live-birth deliveries—an astonishing success rate (in terms of deliveries per retrieval) of 39.8% (Reynolds, *et al.*, 2001).

However, as indicated by the WSJ article, and a wide variety of other clinical studies, including the Reynolds, *et al.*, 2001, study noted above, IVF patients are more likely to have multiple-infant births than women who conceive naturally; furthermore, these multiple pregnancies are known to increase the risks for a broad spectrum of problems, ranging from premature delivery and low birth weight, to such long-term disabilities as cerebral palsy, among surviving babies. For example, Patterson et al, 1993 report that the chance of a twin pregnancy resulting in a baby with cerebral palsy is 8 times that of a singleton birth.

The vast majority of pregnancies with three or more babies are due to IVF and other such “assisted reproductive technologies” (fewer than 20 % arise from natural conception); and such multiple births contribute disproportionately to infant and maternal morbidity and mortality rates, with corresponding increased contributions to health care costs. Consequently, many national and professional organizations in the U.S., Canada and other western countries have provided guidelines on the number of embryos to transfer, in an attempt to balance the desire for success against the risk of multiple births.

The primary objective of this chapter is to examine the fundamental problem of multiple births in IVF from a probabilistic perspective. In what follows, first we review a few representative clinical studies and recommended IVF practice guidelines, and then develop a probability model for IVF and validate it against clinical data. Finally, with the probability model as the basis, we pose — and then solve — the optimization problem of maximizing the chances for success while simultaneously reducing the risk of multiple births. The various consensus qualitative recommendations and guidelines are then interpreted in the context of the probability model and the optimal solution obtained from it.

---

## 11.2 In-Vitro Fertilization and Multiple Births

### 11.2.1 Background and Problem Definition

The primary issue confronting fertility physicians (and patients undergoing IVF treatment) is twofold:

- *Which* embryos should be selected for transfer, and

- *How many* embryos should be transferred.

These questions remain exceedingly difficult to answer because of an intrinsic characteristic of IVF treatment: uncertainty. Whether any particular embryo will implant and develop into pregnancy and ultimately lead to the birth of a healthy child is fundamentally uncertain. The reasons for this fact are many, ranging from “egg factors” such as the source of the egg (donor, self, fresh, cryopreserved, etc), embryo morphology, stimulation protocols, laboratory specifications; and “uterine factors” such as patient age, medical history, and other determinants of uterus status. Thus, in addition to the uncertainties associated with the *implantation potential* of each individual candidate embryo, there are also uncertainties associated with gestation, fetal development and finally childbirth.

While normal fertile couples are not entirely immune to many of these uncertainties, IVF patients, by definition, are particularly prone to poor prognosis so that, from the very start, the chances of successful treatment are typically not very high. To improve the odds of success and ensure an acceptable pregnancy and live birth delivery rate, the prevalent strategy has been to implant multiple embryos. However, this increase in success rate occurs simultaneously with the undesirable consequence of increased risk of multiple pregnancies—along with the associated problems, and consequent implications for health care costs.

Clearly then, the defining problem of modern IVF practice is *how to balance the risks of multiple births judiciously against the desire to increase the chances that each treatment cycle will be successful* — an optimization problem involving the determination, for each patient, in any particular IVF cycle, the optimum number of embryos to transfer to maximize the chances of a singleton live birth while simultaneously reducing the risk of multiple births.

A multitude of clinical studies have been conducted in an attempt to find a practical, implementable solution to this problem; and the essence of the current “state-of-the-art” is captured by the following summary statements which precedes a recent set of guidelines published in 2006 in the *J Obst Gynecol Can*:

“The desired outcome of infertility treatment is the birth of a healthy child. As multifetal gestations are associated with higher rates of morbidity and mortality, their disproportionately high occurrence after IVF-ET [in vitro fertilization-embryo transfer] should be minimized. The transfer of fewer embryos per attempt should be employed as primary prevention. However, indiscriminate application of limitations upon the number of embryos transferred would be inappropriate until accurate predictors of successful implantation can be determined. Decisions on the number of embryos to transfer should be based upon prognosis determined by variables including the woman's age, prior outcomes, and the number and quality of embryos available for transfer, and should

be made to minimize the risk of multifetal gestation while maintaining a high probability of healthy live birth.”<sup>1</sup>

The extent of the results of these studies (reviewed shortly) is a collection of sound recommendations, based on careful analyses of specific clinical data sets to be sure, but no explicit solution to the optimization problem. To the best of our knowledge, to date, there is in fact no systematic, explicit, quantitative solution to the IVF optimization problem whereby the optimum number of embryos to transfer can be prescribed concretely for each individual patient. Such a solution is developed in this chapter.

### 11.2.2 Clinical Studies and Recommended Guidelines

The literature on the topic of “Assisted Reproduction,” even when restricted to papers that focus explicitly on the issue of IVF and multiple births, is quite extensive. This fact alone makes an exhaustive review next to impossible within the context of this chapter. Nevertheless, it is possible to discuss a few key papers that are particularly relevant to the objectives of this chapter (the application of probability models to problems of practical importance).

#### Factors affecting live-birth and multiple-birth rates

The first group of papers, exemplified by Schieve *et al.*, 1999; Engmann, *et al.*, 2001; Reynolds *et al.*, 2001; Jansen, 2003; and Vahrtanian, *et al.*, 2003, use retrospective analyses of various types of clinical IVF data to determine what factors influence live-birth rates and the risk of multiple births. The main conclusions in each of these studies were all consistent and may be summarized as follows:

1. Patient age and the number of embryos transferred independently affected the chances for live birth and multiple birth.
2. In general, live-birth rates increased if more than 2 embryos are transferred.
3. The number of embryos needed to achieve “maximum live birth rates” varied with age. For younger women (age < 35 years), maximum live-birth rates were achieved with only two embryos transferred; for women age > 35 years, live birth rates were lower in general, increasing if more than 2 embryos were transferred.
4. Multiple birth rates generally increased with increased number of embryos transferred but in an age-dependent manner, with younger women (age < 35 years) generally showing higher multiple-birth risks than older women.

---

<sup>1</sup>Guidelines for the Number of Embryos to Transfer Following In Vitro Fertilization, *J Obstet Gynaecol Can* 2006;28 (9)799-813

5. *Special Cases:* For IVF treatments using donor eggs, the age of the donor rather than maternal age was more important as a determinant of the risk of multiple-birth (Reynolds, *et al.* 2001). Also, success rates are lower in general with thawed embryos than with fresh ones (Vahrtian, *et al.*, 2003).

These conclusions, supported concretely by clinical data and rigorous statistical analyses, are, of course, all perfectly in line with common sense.

### **Prescriptive Studies**

The next group, exemplified by Austin, *et al.*, 1996; Templeton and Morris, 1998; Stradel *et al.*, 2000; and Thurin *et al.*, 2004, are more *prescriptive* in that each in its own way sought to provide explicit guidance—also from clinical data—on the number of embryos to transfer in order to limit multiple births. A systematic review in Pandian *et al.*, 2004 specifically compares the effectiveness of elective two-embryo transfer with single-embryo transfer and transfers involving more than 2 embryos. The Thurin *et al.*, 2004, study is unique, being representative of a handful of randomized prospective studies in which, rather than analyze clinical data “after the fact” (as with other studies), they collected their own data (in real-time) after assigning the treatment applied to each patient (single-embryo transfer versus double-embryo transfer) in randomized trials. Nevertheless, even though these studies were all based on different data sets from different clinics, utilized different designs, and employed different methods of analysis, the conclusions were all remarkably consistent:

1. The risk of multiple births increased with increasing number of (good quality) embryos transferred, with patients younger than 40 at higher risk;
2. The rate of multiple births can be reduced significantly by transferring no more than two embryos;
3. By performing single-embryo transfers (in selected cases), the rate of multiple births can be further reduced, although at the expense of reduced rate of live births;
4. Consecutive single-embryo transfers (one fresh embryo transfer followed—in the event of a failure to achieve term pregnancy—by one additional frozen-and-thawed embryo transfer) achieves the same significant reduction possible with single-embryo transfer without lowering the rate of live births substantially below that achievable with a one-time double-embryo transfer.

### Determining Implantation Potential

Central to the results of these “prescriptive” studies is an underlying recognition that, as an aid in rational decision-making, a reasonable estimate of the quality of each transferred embryo, (or equivalently, its “implantation potential”) is indispensable. This is especially so for elective single-embryo transfers where, for obvious reasons, success requires using the best quality embryos with the highest implantation potential.

Quantitatively determining the implantation potential of embryos clearly remains a very difficult proposition, but Bolton, *et al.*, 1989, and Geber and Sampaio, 1999, have proposed techniques for carrying out this task, specifically to facilitate the process of embryo selection for maximizing the success rate of IVF treatment. Thus, while non-trivial, it is indeed possible to determine the clinical chances that the transfer of a single embryo will lead to a livebirth. This fact is important for the model-based analysis that follows shortly.

### Qualitative Optimization Studies

Deciding on the appropriate number of embryos to transfer in each IVF cycle is, as noted earlier, really an optimization problem because the fundamental issue boils down to *maximizing* IVF success rates while simultaneously minimizing the risk of multiple births, objectives that are fundamentally conflicting. Very few studies have taken such an explicitly technical “optimization” perspective of the problem. In Yaron, *et al.*, 1997, for example, the authors present a retrospective study of 254 oocyte-donation IVF patients. Even though the word “optimal” appears in the paper’s title, the text of the paper itself contains no more than a brief discussion at the tail end of a collection of suggestions arising from the data analysis—there is no optimization (formal or informal), and no concrete prescription one way or another in the paper.

In Combelles, *et al.*, 2005, the authors specifically focus on IVF patients older than 40 years of age; and through a retrospective study of data on 863 patients covering a period of more than 5 years the authors arrived at the conclusion that for this group of patients, the optimum number of embryos to transfer is 5. The finding was based on the following results from their data analysis:

1. Transferring 5 or more embryos resulted in significantly increased pregnancy and live birth rates compared with transferring fewer than 5;
2. Transferring more than 5 embryos did not confer any significant additional clinical outcome.

We revisit these results later.

The retrospective study reported in Elsner *et al.*, 1997, involving a large population of patients, 2,173 in all, from the clinic operated by the authors, includes an extensive qualitative discussion regarding how best to maximize

IVF success rates without unduly increasing the risk of multiple births. The key conclusions of the study are:

1. Ideally, a single embryo transfer would be optimal if the implantation rates (per embryo) were as high as 50%;
2. Embryos should be screened, and only the few with high potential implantation rates should be selected for transfer.
3. No more than two embryos should be transferred per attempt. To offset the potential lowering of the IVF success rate, the rest should be cryopreserved for subsequent frozen-thaw embryo transfer.

Because it is comprehensive and detailed, but especially because its presentation is particularly appropriate, the results from this study are used to validate the model presented in the next section.

The final category to discuss are the guidelines and policy recommendations developed by professional organizations and various governmental agencies in western nations particularly the US, Canada, the United Kingdom, and Sweden. For example, in 1991, the British Human Fertilisation and Embryology Authority (HFEA) imposed a legal restriction on the number of allowable embryos transferred to a maximum of 3; Sweden, in 1993 recommended a further (voluntary) reduction in the number of embryos transferred from 3 to 2. The American Society of Reproductive Medicine recommended in 1999<sup>2</sup> that no more than two embryos should be transferred for women under the age of 35 who produce healthy embryos; three for those producing poor embryos. A further tightening of these (voluntary) recommendations in 2004 now suggests that women younger than 35 years old with good prognoses consider single-embryo transfers with no more than two embryos only under "extraordinary circumstances". For women aged 35-37 years the recommendation is two embryos for those with good prognoses and no more than 3 for those with poorer prognoses. The Canadian guidelines issued in 2006 referred to earlier in section 11.2.1 are similar, but more detailed and specific. Because they are consistent with, and essentially capture and consolidate all the results of the previously highlighted studies into a single set of cohesive points, the key aspects are presented below:

1. Individual IVF-ET (embryo transfer) programs should evaluate their own data to identify patient-specific, embryo-specific, and cycle-specific determinants of implantation and live birth in order to develop embryo transfer policies that minimize the occurrence of multifetal gestation while maintaining acceptable overall pregnancy and live birth rates.
2. In women under the age of 35 years, no more than two embryos should

---

<sup>2</sup>American Society of Reproductive Medicine. Guidelines on number of embryos transferred. Birmingham, Alabama, 1999

be transferred in a fresh IVF-ET cycle. In women under the age of 35 years with excellent prognoses, the transfer of a single embryo should be considered.

3. In women aged 35 to 37 years, no more than three embryos should be transferred in a fresh IVF-ET cycle. In those with high-quality embryos and favorable prognoses, consideration should be given to the transfer of one or two embryos in the first or second cycle.
4. In women aged 38 to 39 years, no more than three embryos should be transferred in a fresh IVF-ET cycle. In those with high-quality embryos and favorable prognoses, consideration should be given to the transfer of two embryos in the first or second cycle.
5. In women over the age of 39 years, no more than four embryos should be transferred in a fresh IVF-ET cycle.
6. In exceptional cases when women with poor prognoses have had multiple failed fresh IVF-ET cycles, consideration may be given to the transfer of more embryos than recommended above in subsequent fresh IVF-ET cycles.

We now develop a theoretical probability model of IVF, and validate it against the Elsner *et al*, clinical data. The validated model is then used to provide an explicit quantitative expression for determining the theoretical optimum number of eggs to implant. Finally the model results are compared to those from the just-reviewed clinical studies.

---

### 11.3 Probability Modeling and Analysis

#### 11.3.1 Model Postulate

Consider the following central characteristics of the IVF process:

- The fertilization and transfer of each embryo ultimately either results in a live birth (considered a “success”), or not;
- Which of these two mutually exclusive outcomes is the final result of the transfer of a single embryo is uncertain, with such factors as the nature and quality of the embryo, the patient age, other indicators of uterine condition, etc, jointly affecting the outcome in ways not easily quantified explicitly (at least with currently available technology);
- The transfer of  $n$  embryos at once in one IVF treatment cycle is tantamount to  $n$  simultaneous attempts, with the primary objective of im-

proving the chances that at least one embryo will implant and lead to a live birth;

- How many (and which ones) of the  $n$  transferred embryos will ultimately lead to live births is also uncertain.

If the transfer of  $n$  embryos can be considered as “ $n$  independent (Bernoulli) trials under identical conditions;” and if the overall effect of the collection of factors that influence the ultimate outcome of each single “trial”—the transfer of a single embryo—is captured in the parameter  $p$  representing the *probability that a particular single embryo will lead to a successful pregnancy*; then observe that  $X$ , the number of live births in a delivered pregnancy following an IVF treatment cycle involving the transfer of  $n$  embryos, is a binomial random variable whose pdf is as given in Chapter 8, i.e.:

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad (11.1)$$

an expression of the probability of obtaining  $x$  live-born babies from  $n$  embryos. When  $x = 1$ , the live birth is said to result in a “singleton”—the most desirable outcome; a “multiple-birth” is said to occur when  $x = 2$  (fraternal twins), or 3 (triplets), or 4 (quadruplets), . . . , etc. up to and including  $n$ . How this postulated model matches up with real clinical data is examined shortly.

The characteristics of the binomial random variable and its model have been discussed in Chapter 8 and the reader may wish to pause at this point to review these. Within the context of IVF, the parameter  $p$  has a very specific physiological interpretation: it is what is referred to in Jansen, 2003, as “a woman’s total chance for a live birth from one retrieval.” We will refer to it in the rest of this chapter as the “single embryo probability of success” (or SEPS) parameter. It is sometimes referred to as “the embryo implantation potential” in the ART literature, indicative of its characteristic as a composite of both embryo and uterine properties. If this parameter is known, even approximately (see the discussion to follow about the sensitivity of the model results to the degree of accuracy to which  $p$  is determined), then the mathematical model in Eqn (11.1) allows us to carry out a wide variety of theoretical analyses regarding IVF, including outcome prediction, estimation (patient characterization), and optimization.

### 11.3.2 Prediction

Consider, for example, a case where the combined patient/embryo conditions are characterized by the SEPS parameter  $p = 0.2$  (indicating a 20% chance of “success” for each embryo). The binomial model allows us to say the following about the transfer of  $n = 5$  embryos, for instance:

1. Because  $E(X) = np$  for the binomial random variable  $X$ , in this particular case,  $E(X) = 1$ , implying that the expected outcome of this IVF treatment cycle is 1 live birth;

**TABLE 11.1:** Theoretical distribution of probabilities of possible outcomes of an IVF treatment with 5 embryos transferred and  $p = 0.2$ 

| $x$<br>No. of live<br>births in a<br>delivered pregnancy | $f(x)$<br>Probability<br>of occurrence | $\eta(x)$<br>Expected total no.<br>of patients (out of 1000)<br>with pregnancy outcome $x$ |
|--|--|--|
| 0  | 0.328                                  | 328  |
| 1  | 0.410                                  | 410  |
| 2  | 0.205                                  | 205  |
| 3  | 0.051                                  | 51   |
| 4  | 0.006                                  | 6  |
| 5  | 0.000                                  | 0  |

2. Since the theoretical variance,  $\sigma^2 = np(1 - p) = 0.8$ , (so that the standard deviation,  $\sigma = 0.89$ ), the general implication is that there is a fair amount of variability associated with the expected outcomes in this specific treatment scenario. In fact,
3. The full probability distribution can be computed as shown in Table 11.1, indicating a 32.8% chance that the IVF treatment will not succeed in producing a child, but a somewhat higher 41.0% chance of a singleton; a 20.5% chance of twins, a 5.1% chance of triplets, and less than 1% chance of quadruplets or quintuplets. A common alternative interpretation of the indicated probability distribution is shown in the last column: in a population of 1,000 “identical” patients undergoing the same treatment, under essentially identical conditions, as a result of the transfer of 5 embryos to each patient, 328 patients will produce no live births, 410 will have singletons, 205 will have twins, 51 will have triplets, 6 will have quadruplets, and none will have quintuplets.
4. From this table we see that there is more than a 99% chance that  $0 < x < 3$ , with the following implication: while the *expected* outcome is a singleton, the actual outcome is virtually guaranteed to be anything from a complete failure to a triplet and everything in between; it is highly unlikely to observe any other outcome.

### 11.3.3 Estimation

The practical utility of the binomial model for IVF clearly depends on knowing the lone model parameter  $p$  that characterizes the probability of a single embryo transfer leading to a successful live birth. In the absence of reliable technology for determining an appropriate value directly from physiological measurements, this parameter value must then be determined from clinical data, with best results when the data sets are generated from carefully designed experiments.

Consider, for example, the following statement taken from the WSJ article mentioned earlier:

"In 1999, based on results from over 35,000 IVF treatments, the Centers for Disease Control and Prevention reported that between 10% and 13% of women under 35 who had three embryos introduced got pregnant with triplets."

This statement translates as follows: for the women in this study,  $n = 3$  and  $0.1 < P(X = 3) < 0.13$ . From the binomial model in Eqn (11.1), with an unknown  $p$  and  $n = 3$ , we know that

$$P(X = 3) = f(3) = p^3 \quad (11.2)$$

and upon substituting the limiting values of 0.1 and 0.13 for the probability of obtaining triplets, we immediately obtain

$$\hat{p} = [0.46, 0.51] \quad (11.3)$$

as the corresponding estimates of  $p$ . This assumes, of course, that this group of women is "reasonably" homogeneous in the sense that, while not necessarily identical, the relevant individual physiological characteristics are similar. The women participating in this study are therefore characterized (on average) by  $0.46 < p < 0.51$ , with the implication that for this category of women (under the age of 35) there is a 46-51% chance of a single embryo leading to a live birth, a relatively high IVF success rate.

More generally, one can use clinical data records of the following type: (i) *The patients*: a cohort of  $N_n$  patients, each receiving the same number of transferred embryos,  $n$ ; (ii) *The results*: After the IVF treatment,  $\eta_n(1)$  is the total number of singleton births,  $\eta_n(2)$  the total number of twins, or, in general,  $\eta_n(x)$  is the total number of " $x$ -births" ( $x = 3$  for triplets;  $x = 4$  for quadruplets, etc). Provided that *all the patients in the cohort group are similarly characterized with a common SEPS parameter  $p$* , then — as discussed fully in Part IV — the "maximum likelihood" estimate of  $p$  for the group, (say  $\hat{p}_n$ ) is given by:

$$\hat{p}_n = \frac{\text{Total number of live births}}{\text{Total number of transferred embryos}} \quad (11.4)$$

$$= \frac{\sum_{x=1}^n x\eta_n(x)}{nN_n} \quad (11.5)$$

Thus, for example, one of the entries in the data set found in Table I of Elsner, *et al.*, 1997, indicates that 661 patients each received 3 embryos resulting in 164 singletons, 74 twins, and 10 triplets, with no higher order births. In our notation,  $n = 3$ ,  $N_3 = 661$ ,  $\eta_3(1) = 164$ ,  $\eta_3(2) = 74$ , and  $\eta_3(3) = 10$ , so that the estimate of  $p$  for this cohort is given by:

$$\hat{p} = \frac{164 + 2 \times 74 + 3 \times 10}{3 \times 661} = 0.172 \quad (11.6)$$

Note the very important assumption that *all* 661 patients in this cohort group have “identical” (or at least “essentially similar”) characteristics. We shall have cause to revisit this data set and these implied assumptions in the next section.

## 11.4 Binomial Model Validation

Before proceeding to use the binomial model for IVF optimization, we wish first to validate the model against clinical data available in the literature. Primarily because of how they are reported, the data sets presented in the Elsner, *et al.*, 1997, study (briefly referred to in the previous subsection) are structurally well-suited to the binomial model validation exercise. But this was not a study designed for model validation, otherwise the design would have required more control for extraneous sources of variability within each cohort group. Nevertheless, one can still put these otherwise rich data sets to the best use possible, as we now show.

### 11.4.1 Overview and Study Characteristics

The data sets in question are from a retrospective study of 2,173 patients on which fresh and frozen-thawed embryo transfers were performed in the authors’ own clinic over a 42-month period from September 1991 to March 1995. A total number of 6,601 embryos were transferred ranging from 1 to 6 embryos per transfer. Most importantly, the data are available for cohort groups of  $N_n$  patients, receiving  $n = 1, 2, \dots, 6$  embryos; and on  $\eta_n(x)$ , the number of patients with pregnancy outcome  $x$  ( $x = 1$  for singletons, 2 for twins, 3 for triplets, etc), presented separately for each cohort group, making them structurally ideal for testing the validity of the binomial model. Table 11.2 shows the relevant data arranged appropriately for our purposes (by cohort groups according to embryos received, from 1 through 6).

For each cohort group  $n = 1, 2, 3, 4, 5, 6$ , the estimates of the probability of “success” are obtained from the data as  $\hat{p}_1 = 0.097$ ;  $\hat{p}_2 = 0.163$ ;  $\hat{p}_3 = 0.172$ ;  $\hat{p}_4 = 0.149$ ;  $\hat{p}_5 = 0.111$ ;  $\hat{p}_6 = 0.125$  for an overall probability of success for the entire study group  $\hat{p} = 0.154$ . Some important points to note:

- These values are the same as the “embryo implant” value computed by Elsner, *et al.*;
- Although the overall group average is 0.154, the values for each cohort group range from a low of 0.097 for those receiving a single embryo to a high of 0.172 for those receiving 3 embryos.
- As noted in the paper, the value 0.097 is significantly lower than the

**TABLE 11.2:** Elsner, et al. data of outcomes of a 42-month IVF treatment study

| $x$<br>Delivered<br>pregnancy<br>outcome | No. of patients receiving $n = 1, 2, \dots, 6$ embryos<br>with pregnancy outcome $x$ |             |             |             |             |             | $\eta_T(x)$<br>Total no. patients<br>with pregnancy<br>outcome $x$ |
|--|--|-------------|-------------|-------------|-------------|-------------|--|
|  | $\eta_1(x)$  | $\eta_2(x)$ | $\eta_3(x)$ | $\eta_4(x)$ | $\eta_5(x)$ | $\eta_6(x)$ |  |
| 0  | 205  | 288         | 413         | 503         | 28          | 2           | 1439   |
| 1  | 22   | 97          | 164         | 207         | 13          | 1           | 504  |
| 2  | 0  | 17          | 74          | 84          | 5           | 1           | 181  |
| 3  | 0  | 0           | 10          | 32          | 1           | 0           | 43   |
| 4  | 0  | 0           | 0           | 6           | 0           | 0           | 6  |
| 5  | 0  | 0           | 0           | 0           | 0           | 0           | 0  |
| Total                                    | 227  | 402         | 661         | 832         | 47          | 4           | 2173   |

numbers computed for the 2-, 3-, and 4-, embryo cohort group (which also means that it is significantly lower than the overall group average of 0.154). The implication of this last point therefore is that one cannot assume a uniform value of  $p$  for the entire study involving 6,601 embryos; it also raises the question of whether even the computed  $\hat{p}_i$  for each cohort group can be assumed to be uniform for the entire group (especially groups with large numbers of embryos involved such as the 3- and 4- embryo cohort groups). This issue is addressed directly later.

#### 11.4.2 Binomial Model versus Clinical Data

On the basis of the estimated group probabilities,  $\hat{p}_n$ , the binomial probability model for each group is obtained as in Eq (11.1):

$$f_n(x) = \binom{n}{x} \hat{p}_n^x (1 - \hat{p}_n)^{n-x} \quad (11.7)$$

providing the probability of obtaining pregnancy outcome  $x = 0, 1, 2, \dots, 6$ , for each cohort group receiving  $n$  embryos. Now, given  $N_n$  the number of patients in each cohort group (referred to as the “number of cycles” in the original paper) we can use the model to predict the expected number of patients receiving  $n$  embryos that eventually have  $x = 0, 1, 2, \dots, 6$  as the delivered pregnancy outcome,  $\hat{\eta}_n(x)$ , as follows:

$$\hat{\eta}_n(x) = f_n(x)N_n(x); \quad (11.8)$$

The result is shown in Table 11.3, with a graph comparing the model prediction to the data shown in Fig 11.1.

While the model prediction shows reasonable agreement with the overall data, there are noticeable discrepancies, most notably the over-estimation of the number of singletons and the consistent underestimation of the number of multiple births, especially for the two largest cohort groups—those receiving 3 and 4 embryos. The primary source of these discrepancies is the questionable assumption of uniform  $p$  for each cohort group. Is it really realistic, for example, to expect all 832 patients in the cohort group that received 4 embryos (and the  $832 \times 4$  total embryos transferred in this group) to have similar values of  $p$ ? In fact, this question was actually (unintentionally) answered in the study (recall that the objective of the study was really not the determination of implantation potential for cohort groups).

When the data is segregated by age coarsely into just two sets, the “younger” set for patients  $\leq 36$  years old, and the “older” set for patients  $\geq 37$  years old (as was done in Tables II and III of the original paper and summarized here in Table 11.4 for convenience), the wide variability in the values for the “single embryo probability of success” parameter,  $p$  is evident.

There are several important points to note here: First, observe the less

**TABLE 11.3:** Binomial model prediction of Elsner, *et al.* data in Table 11.2

| $x$                         | No. of patients receiving $n = 1, 2, \dots, 6$ embryos<br>with pregnancy outcome $x$ |             |             |             |             |             | $\eta_T(x)$<br>Total no. patients<br>with pregnancy<br>outcome $x$ |
|-----------------------------|--|-------------|-------------|-------------|-------------|-------------|--|
|                             | $\eta_1(x)$  | $\eta_2(x)$ | $\eta_3(x)$ | $\eta_4(x)$ | $\eta_5(x)$ | $\eta_6(x)$ |  |
| Delivered pregnancy outcome |  |             |             |             |             |             |  |
| 0                           | 204.981  | 281.629     | 375.226     | 436.357     | 26.098      | 1.795       | 1326.09  |
| 1                           | 22.019   | 109.691     | 233.836     | 305.603     | 16.293      | 1.539       | 688.98   |
| 2                           | 0  | 10.681      | 48.575      | 80.261      | 4.069       | 0.550       | 144.13   |
| 3                           | 0  | 0           | 3.363       | 9.369       | 0.508       | 0.105       | 13.34  |
| 4                           | 0  | 0           | 0           | 0.410       | 0.032       | 0           | 0.45   |
| 5                           | 0  | 0           | 0           | 0           | 0           | 0           | 0  |
| Total                       | 227  | 402         | 661         | 832         | 47          | 4           | 2173   |

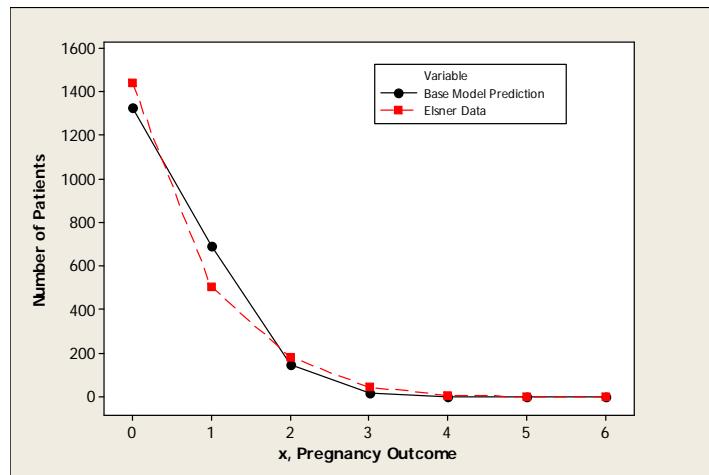


FIGURE 11.1: Elsner data versus binomial model prediction

**TABLE 11.4:** Elsner data stratified by age indicating variability in the “probability of success” estimates

| Embryos<br>recd. ( $n$ ) | Younger ( $\leq 36$ yrs) |                | Older ( $\leq 37$ yrs) |                | Overall |                |
|--------------------------|--------------------------|----------------|------------------------|----------------|---------|----------------|
|                          | Number                   | $\hat{p}$ est. | Number                 | $\hat{p}$ est. | Number  | $\hat{p}$ est. |
| 1                        | 131                      | 0.145          | 96                     | 0.031          | 227     | 0.097          |
| 2                        | 246                      | 0.211          | 156                    | 0.087          | 402     | 0.163          |
| 3                        | 432                      | 0.184          | 229                    | 0.150          | 661     | 0.172          |
| 4                        | 522                      | 0.160          | 310                    | 0.128          | 832     | 0.149          |
| 5                        | 26                       | 0.131          | 21                     | 0.086          | 47      | 0.111          |
| 6                        | 2                        | 0.083          | 2                      | 0.167          | 4       | 0.125          |

obvious fact that for each cohort group,  $n = 1, 2, \dots, 6$ , the overall  $\hat{p}$  estimate is naturally a weighted average of the values estimated for each sub-group (“younger” and “older”); and as one would naturally expect, the weight in each case is the fractional contribution from each sub-group to the total number. Second, and more obvious, is how widely variable the estimates of  $\hat{p}$  are across each cohort group: for example, for the group receiving  $n = 2$  embryos,  $0.087 < \hat{p} < 0.211$ , with the “combined” group value of 0.163 almost twice the value estimated for the “older” sub-group). This latter observation underscores a very important point regarding the use of this particular data set for our model validation exercise: within the context of IVF, the binomial model is an *individual patient* model that predicts the probabilities of various pregnancy outcomes for a specific patient given her characteristic parameter,  $p$ . However, such a parameter—at least in light of currently available technology—can only be estimated from clinical data collected from many patients. Obtaining reasonable estimates therefore requires carefully designed studies involving only patients with a “reasonable expectation” of having similar characteristics. Unfortunately, even though comprehensive and with just the right kind of detail required for our purposes here, the Elsner data sets come from a retrospective study; it is therefore not surprising if many patients in the same cohort group do in fact have different implantation potential characteristics.

One way to account for such non-uniform “within-group” characteristics is, of course, to repeat the modeling exercise for each data set separately using age-appropriate estimates of  $p$  for each cohort sub-group. The results of such an exercise are shown in Figs 11.2 and 11.3. While Fig 11.3 shows a marked improvement in the agreement between the model and the “older” sub-group data, the similarity of the model-data fit in Fig 11.2 to that in Fig 11.1 indicates that even after such stratification by age, significant non-uniformities still exist.

There are many valid reasons to expect significant non-uniformities to persist in the “younger” sub-group: (i) virtually all clinical studies on the effect of age on IVF outcomes (e.g., Schieve *et al.*, 1999; Jansen, 2003; and Vahrtian, *et al.*, 2003,) recognize the age group  $< 29$  years to be different in characteristics from the 29–35 years age group. Even for the “older” sub-group, it is customary to treat the 40–44 years group differently. The data set could thus use a further stratification to improve sub-group uniformity. Unfortunately, only the broad binary “younger”/“older” stratification is available in the Elsner *et al.*, data set. Nevertheless, to illustrate the effect of just one more level of stratification, consider the following postulates:

1. The  $n = 3$  cohort group of 661 patients already stratified into the “younger” 432 ( $\hat{p} = 0.184$ ), and “older” 229 ( $\hat{p} = 0.150$ ) is further stratified as follows: the “younger” 432 separated into 288 with  $\hat{p} = 0.100$  and 144 with  $\hat{p} = 0.352$  (maintaining the original weighted average value of  $\hat{p} = 0.184$ ); and the “older” 229 divided into 153 with  $\hat{p} = 0.100$  and the remaining 76 with  $\hat{p} = 0.25$ , (also maintaining the same original weighted average value of  $\hat{p} = 0.150$ );

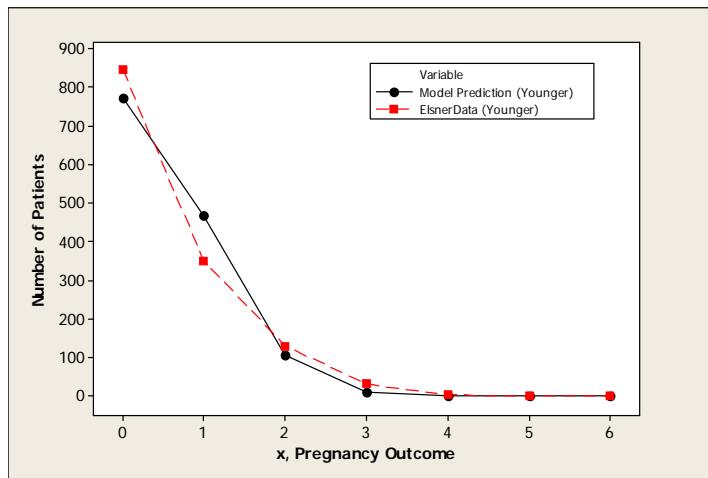


FIGURE 11.2: Elsner data ("Younger" set) versus binomial model prediction

2. The  $n = 4$  cohort group of 832 patients, already stratified into the "younger" 522 ( $\hat{p} = 0.160$ ), and "older" 310 ( $\hat{p} = 0.128$ ) is further stratified as follows: the "younger" 522 separated into 348 with  $\hat{p} = 0.08$  and 174 with  $\hat{p} = 0.320$ ; and the "older" 310 into 207 with  $\hat{p} = 0.08$  and the remaining 103 with  $\hat{p} = 0.224$  (in each case maintaining the original respective weighted average values);
3. The  $n = 5$  cohort group of 47 patients, with only the "younger" 26 with  $\hat{p} = 0.131$  group separated into 17 with  $\hat{p} = 0.06$  and the remaining 9 with  $\hat{p} = 0.265$  (again maintaining the original weighted average value of  $\hat{p} = 0.131$ ).

Upon using this simple stratification of the Elsner data, the results of the stratified model compared with the data is shown first in tabular form in Table 11.5 and in Figs 11.4, 11.5 respectively for the stratified "younger" and "older" data, and Fig 11.6 for the consolidated data. The agreement between the (stratified) model and data is quite remarkable, especially in light of all the possible sources of deviation of the clinical data from the ideal binomial random variable characteristics.

The final conclusion therefore is as follows: given appropriate parameter estimates for the clinical patient population (even very approximate estimates obtained from non-homogenous subgroups), the binomial model matched the clinical data quite well. Of course, as indicated by the parameter estimates in Table 11.4, the value of  $p$  for the patient population in the study is *not* constant but is itself a random variable. This introduces an additional component to the issue of model validation. The strategy of data stratifications by  $\hat{p}$  that we have employed here really constitutes a "manual" (and *ad-hoc*) attempt at dealing

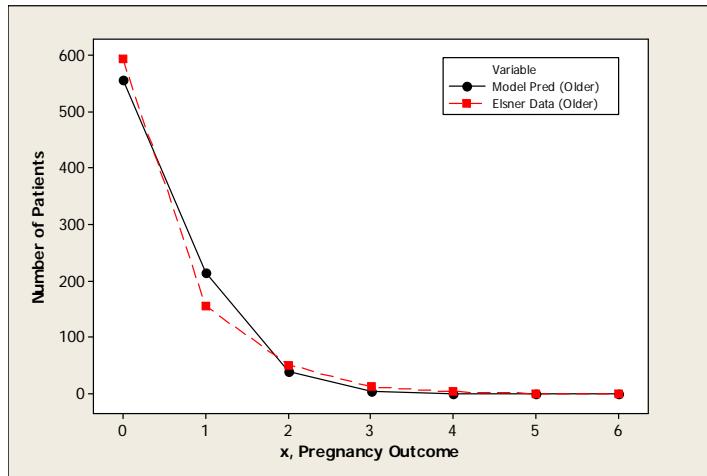


FIGURE 11.3: Elsner data ("Older" set) versus binomial model prediction

**TABLE 11.5:** Stratified binomial model prediction of Elsner, *et al.* data.

| Delivered<br>pregnancy<br>outcome<br>$x$ | Total number of patients<br>with pregnancy outcome $x$ |               |                        |               |         |             |
|--|--|---------------|------------------------|---------------|---------|-------------|
|  | Younger ( $\leq 36$ yrs)                               |               | Older ( $\geq 37$ yrs) |               | Overall |             |
|  | Data   | $\eta_T^y(x)$ | Data                   | $\eta_T^o(x)$ | Data    | $\eta_T(x)$ |
| 0  | 846  | 816           | 593                    | 566           | 1439    | 1382        |
| 1  | 349  | 399           | 155                    | 199           | 504     | 598         |
| 2  | 130  | 118           | 51                     | 43            | 181     | 161         |
| 3  | 31   | 24            | 12                     | 6             | 43      | 30          |
| 4  | 3  | 2             | 3                      | 1             | 6       | 3           |
| 5  | 0  | 0             | 0                      | 0             | 0       | 0           |
| Total                                    | 1359   | 1359          | 814                    | 814           | 2173    | 2173        |

with this additional component indirectly. We have opted for this approach here primarily for the sake of simplicity. A more direct (and more advanced) approach to the data analysis will involve postulating an additional probability model for  $p$  itself, which, when combined with the individual patient binomial model will yield a mixture distribution model (as illustrated in Section 9.1.6 of Chapter 9). In this case, the appropriate model for  $p$  is the Beta distribution; and the resulting mixture model will be the Beta-Binomial model (See Exercise 9.28 at the end of Chapter 9). Such a Beta-Binomial model analysis of the Elsner data is offered as a *Project Assignment* at the end of the chapter. Finally, it is important to note that none of this invalidates the binomial model; on the contrary, it reinforces the fact that the binomial model is a *single patient model*, so that for the mixed population involved in the Elser

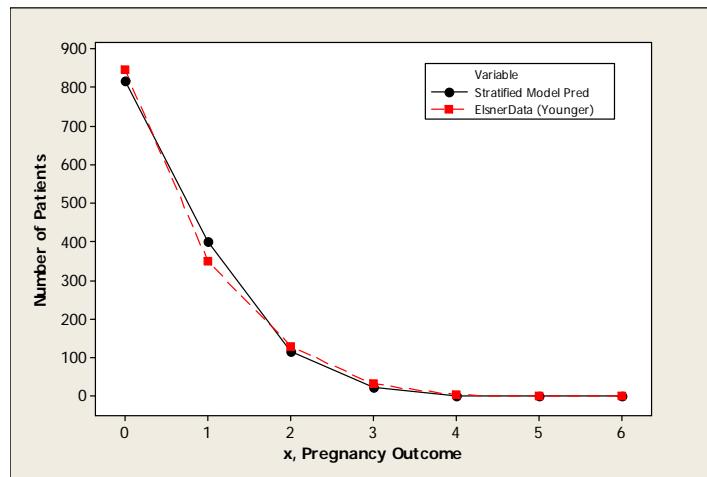


FIGURE 11.4: Elsner data ("Younger" set) versus stratified binomial model prediction

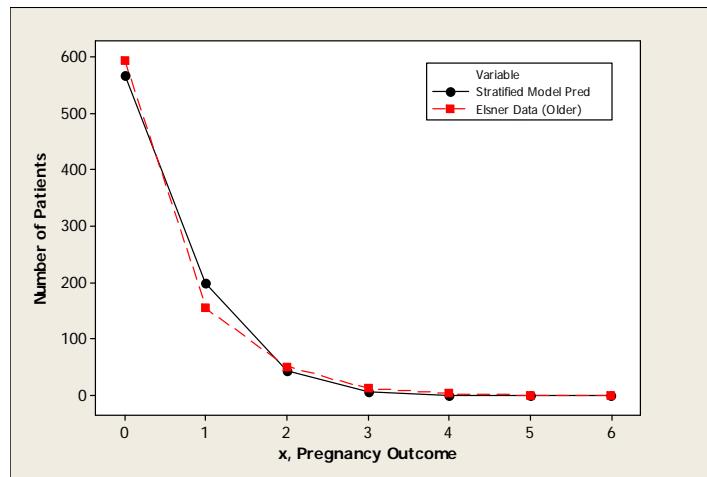


FIGURE 11.5: Elsner data ("Older" set) versus stratified binomial model prediction

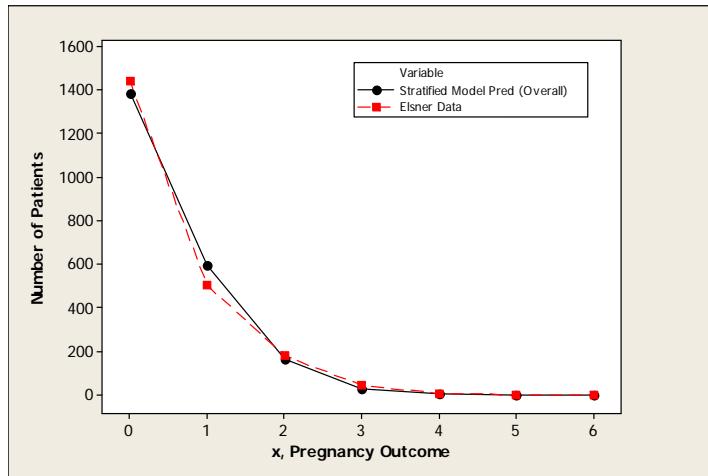


FIGURE 11.6: Complete Elsner data versus stratified binomial model prediction

*et al.* clinical study, the value of  $p$  is better modeled with a pdf of its own, to capture how  $p$  itself is distributed in the population explicitly.

We will now proceed to use the binomial model for analysis and optimization.

## 11.5 Problem Solution: Model-based IVF Optimization and Analysis

For any specified  $p$ , the binomial model provides a quantitative means of analyzing how the probability of each pregnancy outcome,  $x$ , varies with  $n$ , the number of embryos transferred at each treatment cycle. In particular, we are concerned with the following three probabilities:

1.  $P_0 = P(X = 0)$ , the probability of an unsuccessful treatment cycle that produces no live birth;
2.  $P_1 = P(X = 1)$ , the probability of obtaining a singleton (the most desirable pregnancy outcome); and,
3.  $P_{MB} = P(X > 1)$ , the probability of obtaining multiple births (where there is no particular distinction in terms of the actual number once it is greater than 1).

Our interest in these specific probabilities should be obvious: at each IVF cycle, the objective is to reduce the first and last probabilities while maximizing

the second. From the binomial model, these probabilities are given explicitly as:

$$P_0 = (1-p)^n \quad (11.9)$$

$$P_1 = np(1-p)^{n-1} \quad (11.10)$$

$$\begin{aligned} P_{MB} &= P(X > 1) = 1 - (P_0 + P_1) \\ &= 1 - (1-p)^n - np(1-p)^{n-1} \end{aligned} \quad (11.11)$$

Note that these three probabilities are constrained to satisfy the expression:

$$1 = P_0 + P_1 + P_{MB} \quad (11.12)$$

with the all-important implication that any one of these probabilities increases (or decreases) at the expense of the others. Still, each probability varies with  $n$  in a distinctive manner that can be exploited for IVF optimization, as we now show.

### 11.5.1 Optimization

In the most general sense, the “optimum” number of embryos to transfer in any IVF cycle is that number  $n^*$  which simultaneously minimizes  $P_0$ , maximizes  $P_1$ , and also minimizes  $P_{MB}$ . From the model equations, however, observe that (a)  $P_0$  is a monotonically decreasing function of  $n$ , with no minimum for finite  $n$ ; (b) although not as obvious,  $P_{MB}$  has no minimum because it is a monotonically *increasing* function of  $n$ ; but fortunately (c)  $P_1$  does in fact have a maximum. However, the most important characteristic of these probabilities is the following: by virtue of the constraint in Eq. (11.12), maximizing  $P_1$  also simultaneously minimizes the *combined sum* of the undesirable probabilities ( $P_0 + P_{MB}$ )!

We are therefore faced with the fortunate circumstance that the “IVF optimization” problem can be stated mathematically simply as:

$$n^* = \arg \max_n \{np(1-p)^{n-1}\} \quad (11.13)$$

and that the resulting solution, the optimum number of embryos to transfer which maximizes the probability of obtaining a singleton, also simultaneously minimizes the combined probability of undesirable side effects.

The closed-form solution to this problem is obtained via the usual methods of differential calculus as follows:

Since whatever maximizes

$$f_n(1) = np(1-p)^{n-1} \quad (11.14)$$

also maximizes  $\ln f_n(1)$ , solving

$$\frac{d \ln f_n(1)}{dn} = \frac{1}{n} + \ln(1-p) = 0 \quad (11.15)$$

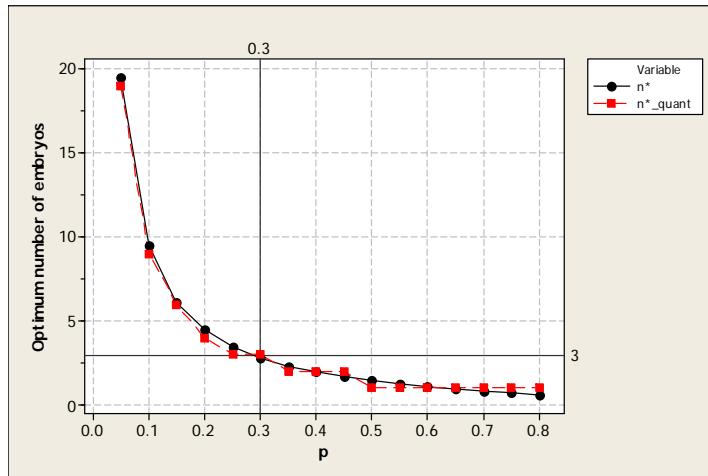


FIGURE 11.7: Optimum number of embryos as a function of  $p$

for  $n$  immediately yields the desired solution,

$$\frac{1}{n^*} = \ln \left( \frac{1}{1-p} \right) \quad (11.16)$$

with the following implications:

Given  $p$ , the probability that a particular single embryo will lead to a successful pregnancy, the optimum number of embryos to transfer during IVF is given by the expression in Eq (11.16) rounded to the nearest integer.

A plot of  $n^*$  as a function of  $p$  is shown in Fig 11.7, the actual continuous value and the corresponding “quantized” (rounded to the nearest integer) value. The indicated reference lines show, as an example, that for patients for whom  $p = 0.3$ , the optimum number of embryos to transfer is 3.

### 11.5.2 Model-based Analysis

The general binomial pdf in Eq (11.1), or the more specific expressions for the probabilities of direct interest to IVF treatment derived from it, and shown in Eqs (11.9), (11.10) and (11.11), provide some insight into the probabilistic characteristics of IVF treatment. For the most desirable pregnancy outcome,

$x = 1$ , the singleton live birth, Fig 11.8 shows the complete surface plot of

$$f_n(1) = np(1 - p)^{n-1} \quad (11.17)$$

as a function of  $n$  and  $p$ . Note the general nature of the surface but in particular the distinctive ridge formed by the maxima of this function. The following are some important characteristics of IVF this figure reveals:

As indicated by the lines sweeping from left to right in the figure, for any given  $n$  embryos transferred, there is a corresponding patient SEPS parameter  $p$  for which the probability of obtaining a singleton is maximized. Furthermore, as  $n$  increases (from back to front), the appropriate “maximal”  $p$  is seen to decrease, indicating that transferring small numbers of embryos works best only for patients with high probabilities of success, while transferring large numbers of embryos works best for patients for whom the probability of success is relatively low. It also shows that for those patients with relatively high probabilities of success (for example, young patients under 35 years of age), transferring large numbers of embryos is counterproductive: the probability of obtaining singletons in these cases is remarkably low across the board (see the flat surface in the bottom right hand corner of the figure) because the conditions overwhelmingly favor multiple births over singletons. All this, of course, is in perfect keeping with current thought and practice; but what is provided by Eq (11.17) and Fig 11.8 is quantitative.

The complementary observation from the lines sweeping from back to front is that for a given single embryo probability of success,  $p$ , there is a corresponding number of embryos to transfer that will maximize the probability of obtaining a singleton. Also, as  $p$  increases (from left to right), this optimum number is seen to decrease. This, of course, is what was shown earlier quite precisely in Fig 11.7.

Finally, when the optimum number of embryos,  $n^*$ , are transferred for each appropriate value of the SEPS parameter,  $p$ , (the mathematical equivalent of walking along the ridge of the mountain-like surface in the figure), the corresponding theoretical maximum probability of obtaining a singleton increases with  $p$  in a manner shown explicitly in Fig 11.9. The indicated “elbow” discontinuity occurring at  $p = 0.5$  is due to the fact that for  $p < 0.5$ ,  $n^* > 1$  and  $f_n^*(1)$  involves integer powers of  $p$ ; but for  $p \geq 0.5$ ,  $n^* = 1$  so that  $f_n^*(1) = p$ , a straight line with slope 1.

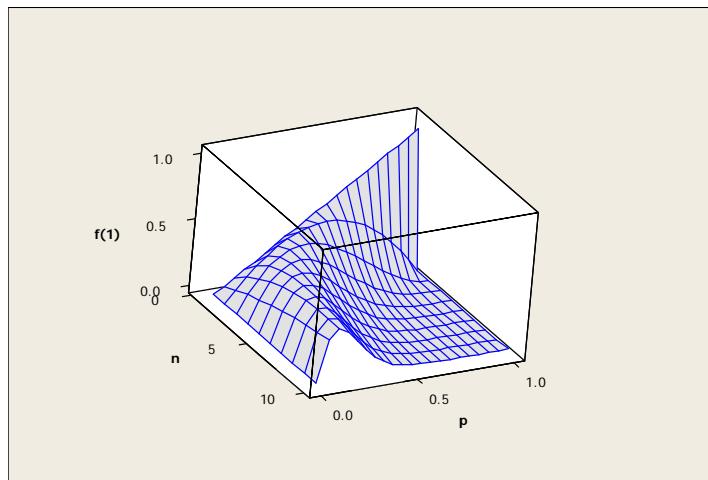
For the sake of completeness, Figures 11.10 and 11.11 show the surface plots respectively for

$$f_n(0) = (1 - p)^n \quad (11.18)$$

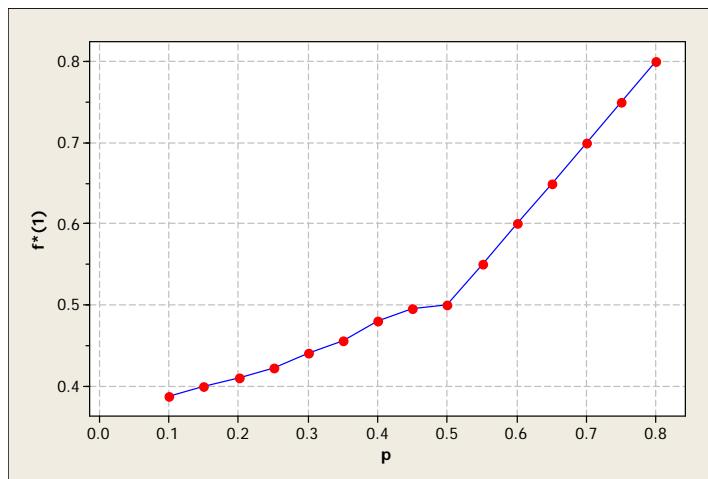
and

$$f_n(m) = 1 - (1 - p)^n - np(1 - p)^{n-1} \quad (11.19)$$

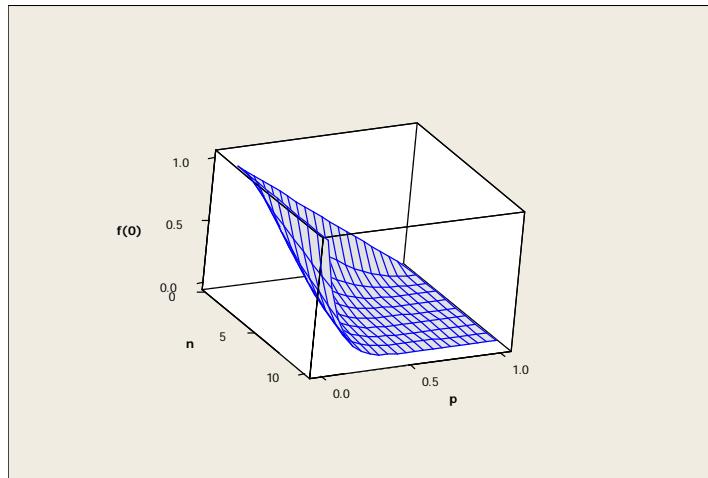
corresponding to Fig 11.8.



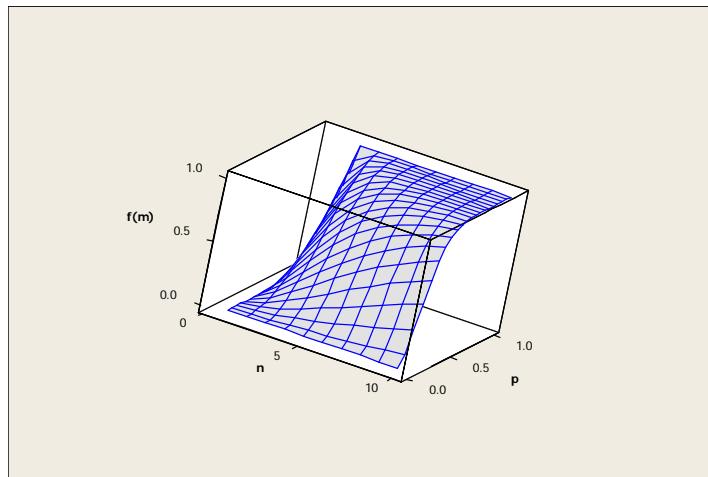
**FIGURE 11.8:** Surface plot of the probability of a singleton as a function of  $p$  and the number of embryos transferred,  $n$



**FIGURE 11.9:** The (maximized) probability of a singleton as a function of  $p$  when the optimum integer number of embryos are transferred



**FIGURE 11.10:** Surface plot of the probability of no live birth as a function of  $p$  and the number of embryos transferred,  $n$



**FIGURE 11.11:** Surface plot of the probability of multiple births as a function of  $p$  and the number of embryos transferred,  $n$

### 11.5.3 Patient Categorization and Theoretical Analysis of Treatment Outcomes

We now return to Fig 11.7 and note that it allows us to categorize IVF patients on the basis of  $p$  (and, by extension, the optimum prescribed number of embryos to transfer) as follows.

1. “Good prognosis” patients:  $p \geq 0.5$ .

For this category of patients,  $n^* = 1$ , with the probability of obtaining a singleton,  $f^*(1) = 0.5$ .

2. “Medium prognosis” patients:  $0.25 \leq p < 0.5$ .

For this category of patients,  $n^* = 2$  or 3, with the probability of obtaining a singleton,  $0.42 < f^*(1) < 0.5$ .

3. “Poor prognosis” patients:  $0.15 \leq p < 0.25$ .

For this category of patients,  $4 < n^* < 6$ , with the probability of obtaining a singleton,  $0.40 < f^*(1) < 0.42$ .

4. “Exceptionally poor prognosis” patients:  $p < 0.15$

For this category of patients,  $n^* > 6$ , but even then the probability of obtaining a singleton,  $f^*(1) \sim 0.40$ .

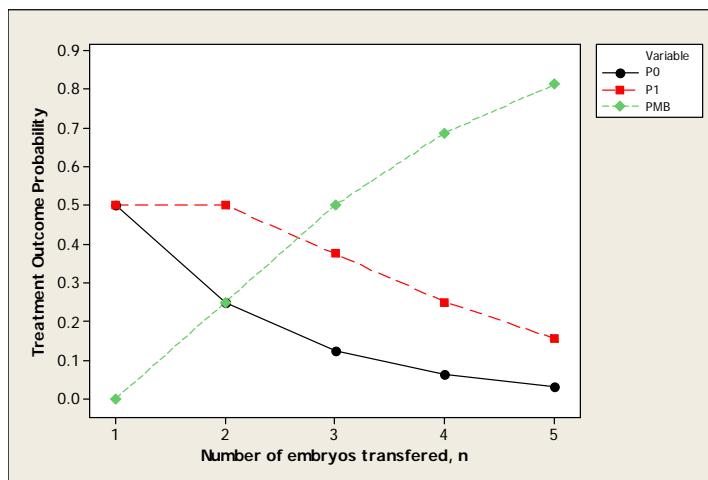
Let us now use the probability model to examine, for each patient category, how the number of embryos transferred influences the potential treatment outcomes.

Beginning with a representative value of  $p = 0.5$  for the “good prognosis” category (e.g. women under age 35, as was the case in the study quoted in the WSJ article that we used to illustrate the estimation of  $p$  in Eqs (11.2) and (11.3)), Fig 11.12 shows a plot of the probabilities of the outcomes of interest,  $P_0$ ,  $P_1$ , and  $P_{MB}$ , as a function of  $n$ .

A few points are worth noting in this figure: first,  $P_1$ , the probability of obtaining a singleton, is maximized for  $n = 1$ , as noted previously; however, this figure also shows that the same probability  $P_1 = 0.5$  is obtained for  $n = 2$ . Why then is  $n = 1$  recommended and not  $n = 2$ ? Because with  $n = 1$ , there is absolutely no risk whatsoever of multiple births, whereas with  $n = 2$ , the probability of multiple births (only twins in this case) is no longer zero, but a hard-to-ignore 0.25.

Note also that transferring more than 1 or 2 embryos for this class of patients actually results in a *reduction* in the probability of a singleton outcome, at the expense of rapidly increasing the probability of multiple births. (Recall the constraint relationship between the probabilities of pregnancy outcomes shown in Eq (11.12).)

When 5 embryos are transferred, there is an astonishing 85% chance of multiple births, and specifically, a 1 in 32 chance of quintuplets (details not shown but easily computed from Eq (11.1)). Thus, going back to the WSJ

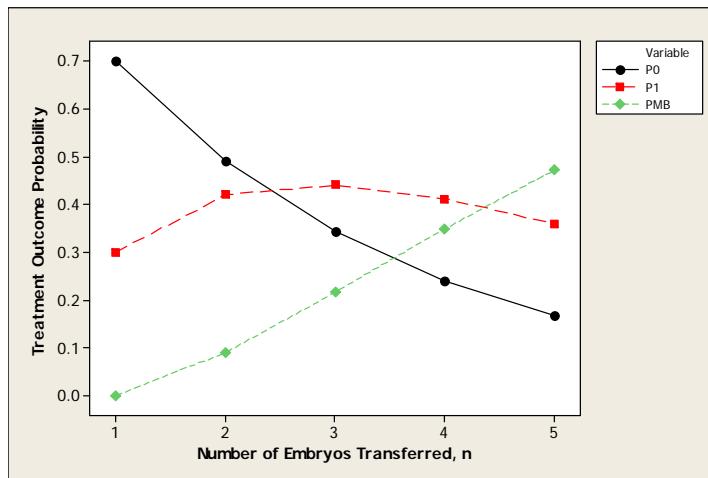


**FIGURE 11.12:** IVF treatment outcome probabilities for “good prognosis” patients with  $p = 0.5$ , as a function of  $n$ , the number of embryos transferred

story of Theresa Anderson with which we opened this chapter, it is now clear that perhaps what her doctor meant to say was that there was a one chance in 30 that *all five embryos would take* rather than that only one would take. With the binomial probability model, one could have predicted the distinct possibility of this particular patient delivering quintuplets because she belongs to the category of patients with good prognosis, for which  $p \geq 0.5$ .

Next, consider a representative value of  $p = 0.3$  for “medium prognosis” patients, which yields the plots shown in Fig 11.13 for the probabilities  $P_0$ ,  $P_1$ , and  $P_{MB}$  as a function of  $n$ . Observe that as noted previously, the optimum  $n$  corresponding to this specific value of  $p = 0.3$  is clearly 3. Transferring fewer embryos is characterized by much higher values for  $P_0$ , the probability of producing no live birth; transferring 4 embryos increases the probability of multiple births more than is offset by the simultaneous reduction in  $P_0$ ; and with 5 embryos, the probability of multiple births dominates all other outcome probabilities.

Finally, when a representative value of  $p = 0.18$  is selected for “poor prognosis” patients, the resulting outcome probability plots are shown in Fig 11.14. First note that for this value of  $p$ , the optimum  $n$  is 5. Recall that the Combelles *et al.*, 2005, study concluded from evidence in their clinical data that  $n = 5$  is “optimum” for women more than 40 years of age. In light of our theoretical analysis, the implication is that for the class of patients referred to in this study,  $p = 0.18$  is a reasonable characteristic parameter. As an independent corroboration of the model-based analysis shown in this figure, consider the following result from the Schieve, *et al.*, 1999, study which states:



**FIGURE 11.13:** IVF treatment outcome probabilities for “medium prognosis” patients with  $p = 0.3$ , as a function of  $n$ , the number of embryos transferred

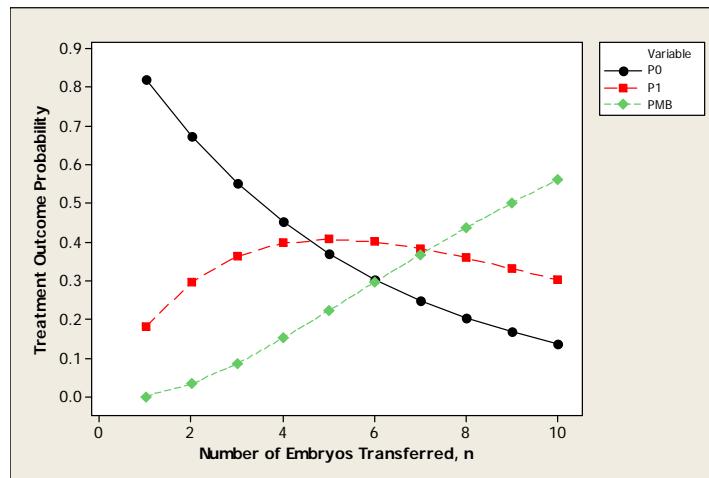
“Among women 40 to 44 years of age, the multiple-birth rate was less than 25% even if 5 embryos were transferred.”

If the patients in this study can reasonably be expected to have characteristics similar to those in the Combelles study, then from our preceding theoretical analysis,  $p = 0.18$  is a reasonable estimate for them also. From our probability model, the specific value for  $P_{MB}$  when  $n = 5$  for this class of patients is therefore predicted to be 0.22 (dotted line and diamond symbol for  $n = 5$ ), which agrees precisely with the above-noted result of the Schieve *et al.*, study.

## 11.6 Sensitivity Analysis

### 11.6.1 General Discussion

Clearly, the heart of the theoretical model-based analysis presented thus far is the parameter  $p$ . This, of course, is in agreement with clinical practice, where embryo transfer policies are based on what the Canadian guidelines refer to as “patient-specific, embryo-specific, and cycle-specific determinants of implantation and live birth.” From such a model-based perspective, this parameter is therefore the single most important parameter in IVF treatment: it determines the optimum number of embryos to transfer, and, in conjunction with the actual number of embryos transferred (optimum or not), determines the various possible pregnancy outcomes and the chances of each one occurring.



**FIGURE 11.14:** IVF treatment outcome probabilities for “poor prognosis” patients with  $p = 0.18$ , as a function of  $n$ , the number of embryos transferred

Given such importance, and since no parameter can be estimated perfectly, it is entirely reasonable to ask: what happens when the model parameter  $p$  is estimated inaccurately, and/or imprecisely? Or, in more practical terms: how sensitive to inevitable parameter estimation errors are the optimization results, model predictions, and other model-based analyses?

It turns out that the binomial model-based analysis of IVF treatment presented thus far is remarkably robust, being quite insensitive to errors in the estimates of  $p$ . But before providing a general, rigorous sensitivity analysis, let us first examine the practical implications of over- or under-estimating  $p$  for some specific cases of practical importance discussed previously. Consider, for example, the “good prognosis” patients for which the true but unknown value of  $p$  is 0.5, so that the true outcome probabilities are as shown in Fig 11.12; overestimating  $p$  as 0.6 (or even 0.7 or higher) leads to a determination of  $n^*$  still as 1 (see Fig 11.7), and the end result will be the transfer of a single embryo, precisely as would have been done in the “nominal” (no error) case, meaning that there are *no* practical consequences for such an overestimation. Conversely, *underestimating p* as 0.4 (a 20% error) leads to an overestimation of  $n^*$  as 2 (again, see Fig 11.7), and from Fig 11.12, we see the primary consequences: the probability of twins increases to 0.25 (from the nominal case of 0) at the same time that the probability of no live birth drops to 0.25 (from the nominal case of 0.5); however, the probability of a singleton remains at the optimum value of 0.5 as in the nominal case.

For “medium prognosis” patients for which  $p = 0.3$  in actual fact but is overestimated as 0.4 (a 33% error), the result is that instead of transferring  $n^* = 3$ , only 2 embryos will be transferred. We see from Fig 11.13 that for the

most desirable outcome probability  $P_1$ , the consequence of transferring only 2 embryos is surprisingly minimal; the more serious consequence is the increase in  $P_0$ , the probability of no live birth, which is somewhat offset by a reduction in  $P_{MB}$  for multiple births. Underestimating  $p$  as 0.25 (a 16.7% error) leads to no change in  $n^*$ ; it will take a 33% underestimation error ( $\hat{p} = 0.2$ ) to change the recommended  $n^*$  to 4. Again, the implications for  $P_1$  is minimal;  $P_0$  is reduced, but at the expense of an increase in  $P_{MB}$ . Still, even under these conditions,  $P_1$  remains the highest of the three probabilities.

For the “poor prognosis” patients characterized in Fig 11.14, with  $p = 0.18$ , overestimating  $p$  as 0.26 or underestimating it as 0.1 (absolute magnitude errors of  $\pm 0.08$ , or  $\pm 44.4\%$ ) leads to the following consequences: overestimation leads to a transfer of 3 embryos instead of 5, with minimal implications for  $P_1$  but with a substantial increase in  $P_0$  that is partially offset by a concomitant reduction in  $P_{MB}$ ; underestimation leads to a seriously asymmetric increase to 9 in the embryos transferred, with serious implications for the balance between the outcome probabilities which shifts in favor of increasing  $P_{MB}$  (while decreasing  $P_0$ ) but at the expense of decreasing the desirable  $P_1$ .

The point of this discussion is that in practical terms, it will take substantial percentage estimation errors (mostly in the direction of underestimating  $p$ ) to affect IVF treatment outcome probabilities to any noticeable extent. The reason for this robustness should be evident from Fig 11.7 which shows how insensitive the value of  $n^*$  is to  $p$  for values in excess of 0.3. Thus, for “good prognosis” patients ( $p > 0.5$ ), almost regardless of the actual specific value estimated for  $p$ ,  $n^*$  is always 1; for “medium prognosis” patients, for  $p$  between 0.25 and 0.5 — a range so broad it covers half the probabilities associated with all patient prognoses not considered as “good” —  $n^*$  is either 2 or 3. Thus the model prescription is completely insensitive for “good prognosis” patients; and its recommendation of 2 or 3 embryos over the wide range  $0.25 \leq p < 0.5$  indicates remarkably mild sensitivity for “medium prognosis” patients. The steep climb in the optimum number of embryos as  $p$  decreases from 0.2 indicates increased sensitivity to errors in  $p$  for “poor prognosis” patients. Nevertheless, given that the nominal values of  $p$  are also low in this range, the relative sensitivity is in fact not as high, as we now show with the following general theoretical derivations.

### 11.6.2 Theoretical Sensitivity Analysis

The general question of interest here is: *how sensitive is the model and its analysis results to errors in  $p$ ?*, or in more practical terms, how good does the estimate of  $p$  have to be so that the prescription of  $n^*$ , and the theoretical analysis following from it, can be considered reliable? Such questions are answered quantitatively with the *relative sensitivity function*, defined in this case as:

$$\mathcal{S}_r = \frac{\partial \ln n^*}{\partial \ln p}. \quad (11.20)$$

It is perhaps best understood in terms of what it means for the transmission of errors in  $p$  into errors in the recommended  $n^*$ : a relative sensitivity,  $\mathcal{S}_r$ , implies that an error  $\Delta p$  in  $p$  translates to an error  $\Delta n^*$  in  $n^*$  according to

$$\frac{\Delta n^*}{n^*} \approx \mathcal{S}_r \frac{\Delta p}{p} \quad (11.21)$$

From the expression in Eq (11.16), we obtain, through the usual calculus techniques, and after some simple algebraic manipulations, the closed form, analytical expression,

$$\mathcal{S}_r = \frac{p}{(1-p) \ln(1-p)} \quad (11.22)$$

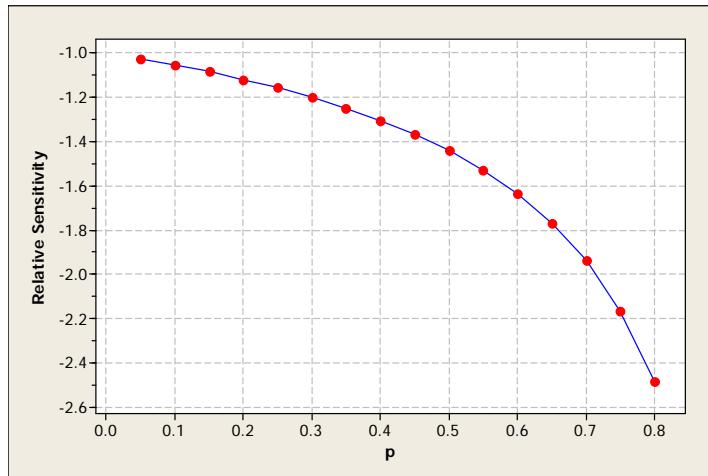
a plot of which is shown in Fig 11.15. First, note in general that  $\mathcal{S}_r$  is always negative and greater than 1 in absolute value. This implies that (i) overestimating  $p$  always translates to underestimating  $n^*$ , and vice versa (as we already alluded to in the preceding general discussion); and (ii) the relative over- or under-estimation error in  $p$  is always magnified in the corresponding relative error in  $n^*$ . Quantitatively, the *specific* information contained in this figure is best illustrated by considering what it indicates about the region  $0 < p < 0.25$  identified in our previous discussion for “poor prognosis” patients as critical in terms of sensitivity to errors in  $p$ . Observe that in this region,  $-1.16 < \mathcal{S}_r < -1.0$ , with the implication that a 10% error in  $p$  estimates translates to no more than an 11.6% error in the recommended  $n^*$ . Keep in mind that in practice, fractional errors in  $n^*$  are inconsequential until they become large enough be rounded up (or down) to the nearest integer. Thus, an 11.6% error on a nominal value of  $n^* = 10$  (or for that matter any error less than 15%) translates to only  $\pm 1$  embryo. Thus, even though from our previous discussion we know that good estimates of  $p$  are required for “poor prognosis” patients, the relative sensitivity function (Eq (11.22), and Fig 11.15) indicate that the model can tolerate as much as 10% error in the estimate of  $p$  with little or no consequence in the final results.

---

## 11.7 Summary and Conclusions

### 11.7.1 Final Wrap-up

The fundamental premise of this chapter is that the random phenomenon inherent to IVF treatment (and attendant issues), fraught as it is with uncertainty, is in fact amenable to systematic analysis via probability theory. In



**FIGURE 11.15:** Relative sensitivity of the binomial model derived  $n^*$  to errors in estimates of  $p$  as a function of  $p$

particular, in an IVF treatment cycle involving the transfer of  $n$  embryos, the pregnancy outcome,  $X$ , is well-modeled as a Binomial,  $Bi(n, p)$  random variable, where the binomial parameter  $p$  represents the single embryo probability of success. Thus, even though the actual pregnancy outcome is uncertain, the probability of obtaining any one of the  $n$  possible outcomes (no live birth, a singleton, twins, triplets, quadruplets, ...,  $n$ -tuplets) is given by the pdf shown in Eq (11.1). Of particular interest are the specific and explicit expressions derived from this general pdf and shown in Eqns (11.9)–(11.11) for the probabilities  $P_0$ ,  $P_1$ , and  $P_{MB}$ , of no live birth, a singleton, and multiple births (all other live births that are *not* *singletons*) respectively.

Even though it was not necessarily designed for validating the binomial model (an exercise that would have required carefully controlling for sources of error due to excessive and undesirable parameter variation, by using only cohort groups with approximately homogeneous characteristics), we were still able to use the clinical data from the Elsner study (Elsner *et al.* 1997) to show that the proposed binomial model indeed provides a reasonable representation of reality.

The primary advantage of such a theoretical model is that it can be used to solve analytically, in a general fashion, the vexing problem of determining the optimum number of embryos to transfer in any particular IVF treatment cycle—an optimization problem requiring the maximization of the chances of delivering a singleton while simultaneously minimizing the chances of obtaining undesirable outcomes (no live births, *and* multiple births). The binomial model provides the additional bonus that a single optimization problem simultaneously achieves both objectives. The final result, the explicit mathematical

expression shown in Eq (11.16) and plotted in Fig 11.7, states this optimum number of embryos,  $n^*$ , explicitly as a function of the parameter  $p$ . Also, we have shown that these results are robust to unavoidable errors in estimating this parameter in practice.

### 11.7.2 Conclusions and Perspectives on Previous Studies and Guidelines

Here, then, is a series of key conclusions from the discussion in this chapter.

First, the following is a list of the characteristics of the binomial model of IVF treatment:

1. It provides a valid mathematical representation of reality;
2. It depends on a single key parameter,  $p$ , whose physiological meaning is clear (the single embryo probability of success—or the embryo implantation potential); and
3. It is used to obtain an explicit expression for the optimum number of embryos to transfer in an IVF cycle, and this result is robust to uncertainty in the single model parameter.

Next, we note that the binomial model-based prescription of the optimum number of embryos to transfer agrees perfectly with earlier heuristics and guidelines developed on the basis of specific empirical studies. While the precise number can be obtained analytically from Eq (11.16), the practical implications of the results may be summarized as follows:

1. For “good prognosis” patients,  $p \geq 0.5$ , transfer only 1 embryo;
2. For “medium prognosis” patients,  $0.25 < p < 0.5$ , transfer 2 embryos for those with  $p \geq 0.35$  and 3 for those with  $p < 0.35$ ;
3. For “poor prognosis” patients,  $p < 0.25$ , transfer  $n > 3$  embryos with the specific value in each case depending on the value of  $p$ , as determined by Eq (11.16) rounded to the next integer: for example,  $n = 4$  for  $p = 0.2$ ;  $n = 5$  for  $p = 0.18$ ,  $n = 6$  for  $p = 0.15$ , etc.

These results agree with, but also generalize, the results of previous clinical studies, some of which were reviewed in earlier sections. Thus, for example, the primary result in the Combelles *et al.*, 2005, study, that  $n = 5$  is optimum for patients older than 40 years, strictly speaking can only be considered valid for the specific patients in the study used for the analysis. The prescription of the binomial model on the other hand is general, and not restricted to any particular “data set;” it asserts that  $n = 5$  is optimal for all patients for which  $p = 0.18$  whether they are 40 years old, younger, or older. This leads to the final set of conclusions having to do with the perspective on previous studies and IVF treatment guidelines provided by the binomial model-based analyses.

In light of the analyses and discussions in this chapter, the most important implication of the demonstrated appropriateness of the binomial model for IVF treatment is that treatment guidelines should be based on the value of the parameter  $p$  for each patient, not so much age. (See recommendation 1 of the Canadian guidelines summarized earlier.) From this perspective, age in the previous studies is seen as a convenient—but not always a perfect—surrogate for this parameter. It is possible, for example, for a younger person to have a lower SEPS parameter  $p$ , for whatever reason, uterine or embryo-related. Conversely, an older patient treated with eggs from a young donor will more than likely have a higher-than-expected SEPS parameter value. In all cases, no conflicts arise if the transfer policy is based on the best estimate of  $p$  rather than age:  $p$  is the more direct determinant of embryo implantation rate; age is an indirect and not necessarily foolproof, indicator.

On the basis of this section's model-based discussion therefore, all the previous studies and guidelines may be consolidated as follows:

1. For each patient, obtain the best estimate of the SEPS parameter,  $p$ ;
2. Use  $p$  to determine  $n^*$  either from the analytical expression in Eq (11.16) rounded to the nearest integer, or else from Fig (11.7);
3. If desired, Eqns (11.9), (11.10) and (11.11) may then be used to analyze outcome probabilities given the choice of the number of embryos to transfer (see for example, Fig 11.13).

Finally, it should not be lost on the reader just how much the probability modeling approach discussed in this chapter has facilitated the analysis and optimization of such a complex and important problem as that posed by IVF outcome optimization. Even with the unavoidable idealization implied in the SEPS parameter,  $p$ , this binomial model parameter provides valuable insight into the fundamental characteristics of the IVF outcome problem. It also allows a consolidation of previous qualitative results and guidelines into a coherent and quantitative three-point guideline enumerated above.

## References

1. Austin, C. M., S.P. Stewart, J. M. Goldfarb, *et al.*, 1996. Limiting multiple pregnancies in in Vitro fertilization/embryo transfer (IVF-ET) cycles, *J. Assisted Reprod and Genetics*, 13 (7) 540-545.
2. Bolton, V.N., S.M. Hawes, C.T., Taylor and J.H. Parsons, 1989. *J In Vitro Fert Embryo Transf.*, 6 (1) 30-35.
3. Combelles, C.M.H., B. Orasanu, E.S. Ginsburg, and C. Racowsky, 2005. Optimum number of embryos to transfer in women more than 40 years of age

undergoing treatment with assisted reproductive technologies, *Fert. and Ster.*, 84 (6) 1637-1642.

4. Elsner, C.W., M.J. Tucker, C.L. Sweitzer, *et al.*, 1997. Multiple pregnancy rate and embryo number transferred during in vitro fertilization, *Am J. Obstet Gynecol.*, 177 (2) 350-357.
5. Engmann, L., N. Maconochie, S.L. Tan, and J. Bekir, 2001. Trends in the incidence of births and multiple births and the factors that determine the probability of multiple birth after IVF treatment, *Hum Reprod.*, 16 (12) 2598-2605.
6. Geber, S. and M. Sampaio, 1999. Blasotmere development after embryo biopsy: a new model to predict embryo development and to select for transfer, *Hum Reprod.*, 14 (3) 782-786.
7. Jansen, R. P. S., 2003. The effect of female age on the likelihood of a live birth from one in-vitro fertilisation treatment, *Med. J. Aust.*, 178, 258-261.
8. Pandian, Z., S. Bhattacharya, O. Ozturk, G.I. Serour, and A. Templeton, 2004. Number of embryos for transfer following in-vitro fertilisation or intra-cytoplasmic sperm injection, *Cochrane Database of Systematic Reviews*, 4, Art No. CD003416. DOI:10.1002/14651858.CDC003416.pub2.
9. Patterson, B., Nelson, K.B., Watson, L. *et al.*, 1993. Twins, triplets, and cerebral palsy in births in Western Australia in the 1980's, *Brit. Med. J.* 307, 1239-1243.
10. Reynolds, M. A., L.A. Schieve, G. Jeng, H.B. Peterson, and L.S. Wilcox, 2001. Risk of multiple birth associated with in vitro fertilization using donor eggs, *Am J. Epidemiology*, 154 (11), 1043-1050.
11. Scheive, L.A., H.B.Peterson, S. Miekle, *et al.*, 1999. Live birth rates and multiple-birth risk using in vitro fertilization, *JAMA*, 282 1832-1838.
12. Strandel, A., C Bergh, and K. Lundin, 2000. Selection of patients suitable for one-embryo transfer may reduce the rate of multiple births by half without impairment of overall birth rates, *Hum Reprod.*, 15 92) 2520-2525.
13. A. Templeton and J. K. Morris, 1998. Reducing the risk of multiple births by transfer of two embryos after in vitro fertilization, *The New Eng J. Med.*, 339 (9) 573-577.
14. A. Thurin, J. Hauske, T Hllensjö, *et al.*, 2004. Elective single-embryo transfer versus double-embryo transfer in in Vitro fertilization, *The New Eng J. Med.*, 351 (23) 2392-2402.
15. A. Vahrtanian, L. A. Schieve, M.A.Reynolds, and G. Jeng, 2003. Live-birth rates and multiple-birth risk of assisted reproductive technology pregnancies concieved using thawed embryos, USA 1999-2000, *Hum Reprod.*, 18 (7), 1442-1448.
16. Yaron, Y, A. Amit, A. Kogosowski, *et al.*, 1997. The optimal number of embryos to be transferred in shared oocyte donation: walking the thin line between low pregnancy rates and multiple pregnancies, *Hum Reprod.*, 12 (4) 699-702

## PROJECT ASSIGNMENT

### Beta-Binomial Model for the Elsner Data.

As noted at the end of Section 11.4, to deal appropriately with the mixed population involved in the Elsner clinical study, a theoretical probability model should be used for  $p$ ; this must then be combined with the binomial model to yield a mixture model. When the Beta  $B(\alpha, \beta)$  model is chosen for  $p$ , the resulting mixture model is known as a *Beta-Binomial* model.

As a project assignment, develop a Beta-Binomial model for the Elsner data in Table 11.2 and compare the model prediction with the data and with the binomial model prediction presented in this chapter.

You may approach this assignment as follows:

The Beta-Binomial mixture distribution arises from a Binomial  $Bi(n, p)$  random variable,  $X$ , whose parameter  $p$ , rather than being constant, has a Beta distribution, i.e., it consists of a conditional distribution,

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x} \quad (11.23)$$

in conjunction with the marginal distribution for  $p$ ,

$$f(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1}; 0 < p < 1; \alpha > 0; \beta > 0 \quad (11.24)$$

1. Obtain the expression for  $f(x)$ , the Beta-Binomial pdf.
2. Show that the mean and variance of this pdf are as follows:

$$E(X) = n \left( \frac{\alpha}{\alpha + \beta} \right) = n\pi \quad (11.25)$$

and, for  $\pi$  defined as in Eq (11.25),

$$Var(X) = n\pi(1 - \pi)\phi; \quad (11.26)$$

with  $\phi$  defined as:

$$\phi = \frac{\alpha + \beta + n}{\alpha + \beta + 1} \quad (11.27)$$

(Caution: These results are not easy to obtain by “brute force.”)

Compare and contrast the expressions in Eqs (11.25) and (11.26) with the corresponding expressions for the Binomial random variable.

3. Compute the mean  $\bar{x}$ , and variance  $s^2$ , of the complete Elsner data shown in Table 11.2. Determine appropriate estimates for the Beta-Binomial model parameters,  $\alpha$  and  $\beta$ , in terms of the values computed for  $\bar{x}$  and  $s^2$  from the data.

4. Plot the theoretical pdf for  $f(p)$  using the values you determined for the parameters  $\alpha$  and  $\beta$ . Discuss what this implies about how the SEPS parameter is distributed in the population involved in the Elsner clinical study.
5. Compute probabilities for  $x = 0, 1, \dots, 6$ , using the Beta-Binomial model and compare with the data.

Write a report describing your model development and data analysis, including comments on the fit of this mixture model to the data versus the binomial model fit that was discussed in this chapter.



— | —

## Part IV

# Statistics

—

|

—

|

—

|

—

|

---

## Part IV: Statistics

*Quantifying Random Variability*

---

*The days of our years are threescore years and ten; and if by reason of strength they be fourscore years, yet is their strength labor and sorrow; for it is soon cut off, and we fly away. . . . So teach us to number our days that we may apply our hearts unto wisdom.*

Moses (c. 1450 BC) *Psalms 90:10,12, KJV*

## Part IV: Statistics

*Quantifying Random Variability*

- **Chapter 12:** Introduction to Statistics
- **Chapter 13:** Sampling
- **Chapter 14:** Estimation
- **Chapter 15:** Hypothesis Testing
- **Chapter 16:** Regression Analysis
- **Chapter 17:** Probability Model Validation
- **Chapter 18:** Nonparametric Methods
- **Chapter 19:** Design of Experiments
- **Chapter 20:** Application Case Studies III: Statistics

# Chapter 12

## *Introduction to Statistics*

|        |   |     |
|--------|---|-----|
| 12.1   | From Probability to Statistics .....                    | 408 |
| 12.1.1 | Random Phenomena and Finite Data Sets .....             | 408 |
| 12.1.2 | Finite Data Sets and Statistical Analysis .....         | 411 |
|        | Populations, Samples and Inference .....                | 412 |
| 12.1.3 | Probability, Statistics and Design of Experiments ..... | 414 |
| 12.1.4 | Statistical Analysis .....                              | 415 |
|        | Descriptive Statistics .....                            | 415 |
|        | Inductive Statistics .....                              | 416 |
|        | Statistical Design of Experiments .....                 | 416 |
| 12.2   | Variable and Data Types .....                           | 417 |
| 12.3   | Graphical Methods of Descriptive Statistics .....       | 418 |
| 12.3.1 | Bar Charts and Pie Charts .....                         | 419 |
| 12.3.2 | Frequency Distributions .....                           | 423 |
| 12.3.3 | Box Plots .....   | 427 |
| 12.3.4 | Scatter Plots .....                                     | 430 |
| 12.4   | Numerical Descriptions .....                            | 435 |
| 12.4.1 | Theoretical Measures of Central Tendency .....          | 436 |
|        | The Mean .....  | 436 |
|        | The Median .....  | 437 |
|        | The Mode .....  | 438 |
| 12.4.2 | Measures of Central Tendency: Sample Equivalents .....  | 438 |
|        | Sample Mean .....                                       | 438 |
|        | Sample Median .....                                     | 439 |
|        | Sample Mode .....                                       | 439 |
| 12.4.3 | Measures of Variability .....                           | 439 |
|        | Range .....   | 440 |
|        | Average Deviation .....                                 | 440 |
|        | Sample Variance and Standard Deviation .....            | 440 |
| 12.4.4 | Supplementing Numerics with Graphics .....              | 442 |
| 12.5   | Summary and Conclusions .....                           | 444 |
|        | APPENDIX .....  | 447 |
|        | REVIEW QUESTIONS .....                                  | 447 |
|        | EXERCISES .....   | 449 |
|        | APPLICATION PROBLEMS .....                              | 454 |

*To understand God's thoughts we must study Statistics  
for these are the measure of His purpose.*

Florence Nightingale (1820–1910)

The uncertainty intrinsic to individual observations of randomly varying phenomena, we now know, need not render mathematical analysis impossible. The discussions in Parts II and III have shown how to carry out such analysis, so long as one is willing to characterize the more stable complete ensemble, and

not the “capricious” individual observations. The key has been identified as the ensemble characterization, via the probability distribution function (pdf),  $f(x)$ , which allows one to carry out probabilistic analysis about the occurrence of various observations of the random variable  $X$ .

In practice, however,  $f(x)$ , the theoretical ensemble model for the random phenomenon in question, is never *fully* available — we may know its form, but the parameters are unknown and must be determined for each specific problem. Only a *finite* collection of individual observations  $\{x_i\}_{i=1}^n$  is available. But any finite set of observations will, like individual observations, be subject to the vagaries of intrinsic variability. Thus, in any and all analysis carried out on this basis, uncertainty is unavoidable. Given this practical reality, how then does one carry out systematic analysis of random phenomena — which requires the full  $f(x)$  — on the basis of finite data? How does one fully characterize the entire ensemble from only a finite collection of observations? Clearly,  $\{x_i\}_{i=1}^n$  is related to, and contains information about, the full  $f(x)$ ; but how are they related, and how can this information be extracted and exploited? What kinds of analyses are possible with finite observations, and how does one cope with the unavoidable uncertainty?

*Statistics* provides rigorous answers to such questions as these by using the very same probability machinery of Parts II and III to deal generally with the theoretical and practical issues associated with analyzing randomly varying phenomena on the basis of finite data. Our study begins in this chapter with an introduction of Statistics first as a conceptual framework complementary to — and dependent on — Probability. We then provide a brief overview of the components of this “Statistical Analysis” framework, as a prelude to the detailed study of each of these components in the remaining chapters of Part IV.

## 12.1 From Probability to Statistics

### 12.1.1 Random Phenomena and Finite Data Sets

A systematic study of randomly varying phenomena involves three distinct but intimately related entities:

1. *X: the actual variable of interest.*

This is the random variable discussed extensively in Chapter 4 and illustrated in Chapter 1 (prior to any formal discussions) with two examples: the yield obtainable from chemical processes (two continuous random variables), and the number of inclusions on a manufactured glass sheet of specified area (a discrete random variable). It is an abstract, conceptual entity.

2.  $\{x_i\}_{i=1}^n$ : *n individual observations; one set out of many other possible realizations of the random variable.*

This is commonly referred to as the “data;” it is the only entity available in practice (from experiments). In the illustrative examples of Chapter 1,  $n = 50$  each for process  $A$ , and process  $B$ ;  $n = 60$  for the glass sheets.

3.  $f(x)$ : *aggregate (or ensemble) description of the random variable; the probability distribution function.*

This is the theoretical model of how the probability of obtaining various results are distributed over the entire range of all possible values observable for  $X$ . It was discussed and characterized generically in Chapter 4, before specific forms were derived for various random variables of interest in Chapters 8 and 9. There we saw that it consists of a functional form,  $f(x)$ , and characteristic parameters; it is therefore more completely represented as  $f(x|\boldsymbol{\theta})$ , which literally reads “ $f(x)$  given  $\boldsymbol{\theta}$ ,” where  $\boldsymbol{\theta}$  is the vector of characteristic parameters. In the first illustrative example of Chapter 1,  $f(x)$  is the Gaussian distribution, with parameters  $\boldsymbol{\theta} = (\mu, \sigma^2)$ ; in the second, it is a Poisson distribution with one characteristic parameter,  $\lambda$ .

Probabilistic random phenomena analysis is based entirely on  $f(x)$ . This is what allows us to abandon the impossible task of predicting an intrinsically unpredictable entity in favor of the more mathematically tractable task of computing the probabilities of observing the randomly varying outcomes. Until now, our discussion about probability and the pdf has been based on the availability of the complete  $f(x)$ , i.e.,  $f(x|\boldsymbol{\theta})$  with  $\boldsymbol{\theta}$  known. This allowed us to focus on the *first* task: computing probabilities and carrying out analysis, *given* any functional form  $f(x)$  with values of the accompanying characteristic parameters,  $\boldsymbol{\theta}$ , assumed known. We have not been particularly concerned with such practical issues as where either the functional form, or the specific characteristic parameter values come from. With the first task complete, we are now in a position to ask important practical questions: in actual practice, what is *really* available about any random variable of interest? and how does one go about obtaining the complete  $f(x)$  required for random phenomena analysis?

It turns out that for any *specific* randomly varying phenomenon of interest, the theoretical description,  $f(x)$ , is never completely available, usually because the characteristic parameters associated with the particular random variable in question are unknown; only finite data in the form of a set of observations  $\{x_i\}_{i=1}^n$ ; ( $n < \infty$ ), is available in practice. The immediate implication is that to apply the theory of random phenomena analysis successfully to practical problems, we must now confront the practical matter of *how* the complete  $f(x)$  is to be determined from finite data, for any particular random variable,  $X$ , of interest. The problem at hand is thus one of *analyzing randomly varying phenomena on the basis of finite data sets*; this is the domain of Statistics. With Probability,  $f(x)$  — the functional form along with the parameters — is

given, and analysis involves determining the probabilities of obtaining specific observations  $\{x_i\}_{i=1}^n$  from the random variable  $X$ . The reverse is the case with Statistics:  $\{x_i\}_{i=1}^n$  is given, and analysis involves determining the appropriate (and complete)  $f(x)$  — the functional form *and* the parameters — for the random variable  $X$  that generated the given specific observation. (Readers familiar with the subject matter of Process Control may recognize that the relationship between Statistics and Probability, even in the skeleton form given above, is directly analogous to the relationship between Process Identification and Process Dynamics, with the classical transfer function model  $g(s)$ , — or any other dynamic model form — playing the role of the pdf  $f(x)$ .) One should not be surprised then that the theoretical concept of the pdf,  $f(x)$ , still plays a significant role in handling this new “reverse” problem. This function provides the theoretical basis for determining, from the finite data set  $\{x_i\}_{i=1}^n$ , the most likely underlying complete  $f(x)$ , and to quantify the associated uncertainty.

It is in this sense that Statistics is referred to as a methodology for *inferring* the characteristics of the complete pdf,  $f(x)$ , from finite data sets  $\{x_i\}_{i=1}^n$ , and for quantifying the associated uncertainty. In general, Statistics is typically defined as a methodology for

1. efficiently extracting information from data; and
2. efficiently quantifying such information,

a definition broad enough to encompass all aspects of the subject matter, as we show in the remaining chapters of Part IV. Thus, while the focus of “Probability” is the random variable,  $X$ , the focus of “Statistics” is the finite data set  $\{x_i\}_{i=1}^n$  as a specific realization of  $X$ .

#### **Example 12.1 PROBABILITY AND STATISTICAL ANALYSIS FOR COIN TOSS EXPERIMENT: A PREVIEW**

Consider the illustrative experiment introduced in Example 3.1 in Chapter 3, and in Example 4.1 of Chapter 4, which involved tossing a coin 3 times and recording the number of observed tails. For a specific coin, after performing this three-coin toss experiment exactly 10 times under identical conditions, the following result set was obtained:  $S_1 = \{0, 1, 3, 2, 2, 1, 0, 1, 2, 2\}$ .

(1) What is the random variable  $X$ , and for a generic coin, what is the theoretical ensemble description  $f(x)$ ? (2) How is the specific experimental result related to the random variable,  $X$ ? (3) For a specific coin for which  $p = 0.5$  compute the probabilities of obtaining 0, 1, 2 or 3 tails in each experiment. How consistent with the observed results are the theoretical probabilities?

#### **Solution:**

(1) Recall from these previous examples and discussion that the random variable  $X$  is *the total number of tails obtained in the 3 tosses*; and from the discussion in Chapter 8, this is a binomial random variable,  $Bi(n, p)$  with  $n = 3$  in this particular case, and  $p$  as the characteristic parameter.

The ensemble description is the binomial pdf:

$$f(x|p) = \binom{3}{x} p^x (1-p)^{3-x}; x = 0, 1, 2, 3 \quad (12.1)$$

from which the probabilities of observing  $x = 0, 1, 2, 3$  can be computed for any given value of  $p$ .

(2) The data set  $S_1$  is an *experimental* realization of the random variable  $X$ ; the variability it displays is characteristic of randomly varying phenomena. The specific values observed are expected to change with each performance of the experiment.

(3) When  $p = 0.5$ , the required probabilities are obtained as  $P(X = 0) = 1/8 = 0.125$ ;  $P(X = 1) = 3/8 = 0.375$ ;  $P(X = 2) = 3/8 = 0.375$ ;  $P(X = 3) = 1/8 = 0.125$ . Strictly from the limited data, observe that two of the 10 values are 0, three values are 1, four values are 2, and one value is 3; the various relative frequencies of occurrence are therefore  $f_r(0) = 0.2$  for the observation  $X = 0$ ;  $f_r(1) = 0.3$ ;  $f_r(2) = 0.4$ ;  $f_r(3) = 0.1$ ; and if one can assume that this data set is “representative” of typical behavior of this random variable, and that the observed relative frequency of occurrence can be considered as an approximate representation of true probability of occurrence, then the observed relative frequency distribution,  $f_r(x)$ , is actually fairly close to the theoretical probabilities computed under the assumption that  $p = 0.5$ . The implication is that the data set appears to be somewhat consistent with the theoretical model when  $p = 0.5$ .

Statistical analysis is concerned with such analysis as illustrated above, but of course with more precision and rigor.

### 12.1.2 Finite Data Sets and Statistical Analysis

Three concepts are central to statistical analysis: Population, Sample, and Statistical Inference.

1. *Population*: This is the complete collection of *all* the data obtainable from a random variable of interest. Clearly, it is impossible to realize the population in actual practice, but as a conceptual entity, it serves a critical purpose: the population is *the full observational “realization”* of the random variable,  $X$ . It is to statistics what the sample space (or the random variable space) is to probability theory — an important statement we shall expand on shortly.
2. *Sample*: A *specific* set of actual observations (measurements) obtained upon performance of an experiment. By definition, this is a (finite) subset of data selected from the population of interest. It is the only information about the random variable that is actually available *in practice*.

3. *Statistical Inference:* Any statement made about the population on the basis of information extracted from a sample. Because the sample will never encompass the entire population, such statements must include a measure of the unavoidable associated uncertainty — a measure of how “reliable” such statements are.

These concepts (especially the first one) require further clarification beyond these brief descriptions.

### Populations, Samples and Inference

Probability theory, with its focus on the random variable,  $X$  (and its uncertain outcomes), utilizes the sample space,  $\Omega$ , and the random variable space,  $V_X$ , as the basis for developing the theoretical ensemble description,  $f(x)$ . In practice, for any specific problem, the focus shifts to the actual observed data,  $\{x_i\}_{i=1}^n$ ; and the equivalent conceptual entity in statistics becomes the population — the observational ensemble to be described and characterized. Now, while the sample space or the random variable space in probability theory can be specified *à-priori* and generically, the population in statistics refers to observational data, making it a specific, *à-posteriori* entity. To illustrate, recall the three-coin toss experiment in Example 12.1. Before any actual experiments are performed, we know  $V_X$ , the random variable space of all possible outcomes, to be:

$$V_X = \{0, 1, 2, 3\}, \quad (12.2)$$

indicating that with each performance of the experiment, the outcome will be one of the four numbers contained in  $V_X$ ; and, while we cannot predict with certainty what any specific outcome will be, we can compute probabilities of observing any one of the four possible alternatives. For the generic coin for which  $p$  is the probability of obtaining a tail in a single toss, we are able to obtain (as illustrated in Example 12.1) an explicit expression for how the outcome probabilities are distributed over the values in  $V_X$  for this binomial random variable. For a *specific* coin for which, say,  $p = 0.5$ , we can compute, from Eq (12.1), the probabilities of obtaining 0, 1, 2, or 3 tails, as we just did in Example 12.1 (also see Table 4.1). In practice, the true value of  $p$  associated with a specific coin is not known *à-priori*; it must be determined from experimental data such as,

$$S_1 = \{0, 1, 3, 2, 2, 1, 0, 1, 2, 2\}, \quad (12.3)$$

as in Example 12.1. This is considered as a single, 10-observation sample for the specific coin in question, one of a theoretically infinite number of other possible samples — a sample drawn from the conceptual population of all such data obtainable from this specific coin characterized by the true, but unknown, value  $p = 0.5$ . Although finite (and hence incomplete as an observational realization of the binomial random variable in question),  $S_1$  contains information about the true value of the characteristic parameter  $p$ .

associated with this particular coin. Determining appropriate estimates of this true value is a major objective of statistical analysis. Because of the finiteness of sample data, a second series of such experiments will yield a *different* set of results, for example,

$$S_2 = \{2, 1, 1, 0, 3, 2, 2, 1, 2, 1\}. \quad (12.4)$$

This is another sample from the *same* population, and a result of inherent variability, we observe that  $S_2$  is different from  $S_1$ . Nevertheless, this new sample also contains information about the unknown characteristic parameter,  $p$ .

Next consider *another* coin, this time, one characterized by  $p = 0.8$ ; and suppose that after performing the three-coin toss experiment say  $n = 12$  times, we obtain:

$$S_3 = \{3, 2, 2, 3, 1, 3, 2, 3, 3, 3, 2, 2\} \quad (12.5)$$

As before, this set of results is considered to be just one of an infinite number of other samples that could potentially be drawn from the population characterized by  $p = 0.8$ ; and, as before, it also contains information about the value of the unknown population parameter.

We may now note the following facts:

1. With probability, the random variable space for this example is finite, specifiable *a-priori*, and remains as given in Eq (12.2) whether  $p = 0.5$ , or 0.8, or any other admissible value.
2. Not so with the population: it is infinite, and its elements depend on the specific value of the characteristic population parameter. Sample  $S_3$  above, for instance, indicates that when  $p = 0.8$ , the population of all possible observations from the three-coin toss experiment will very rarely contain the number 1. (If the probability of observing a tail in a single toss is this high, the number of tails observed in three tosses will very likely consistently exceed 1.) This is very different from what is indicated by  $S_2$  and  $S_1$ , being samples from a population of results observable from tossing a so-called “fair coin” with no particular preference for tails over heads.
3. Information about the true, but unknown, value of the characteristic parameter,  $p$ , associated with each coin’s population is contained in each finite data set.

Let us now translate this illustration to a more practical problem.

Consider an in-vitro fertilization (IVF) treatment cycle involving a 35-year old patient and the transfer of 3 embryos. In this case, the random variable,  $X$ , is the resulting number of live births; and, assuming that each embryo leads either to a single live birth or not (i.e. no identical twins from the same egg) the possible outcomes are 0, 1, 2, or 3, just as in the three-coin toss illustration, with the random variable space,  $V_X$ , as indicated in Eq (12.2). Recall from

Chapter 11 that this  $X$  is also a binomial random variable, implying that the pdf in this case is also as given in Eq (12.1).

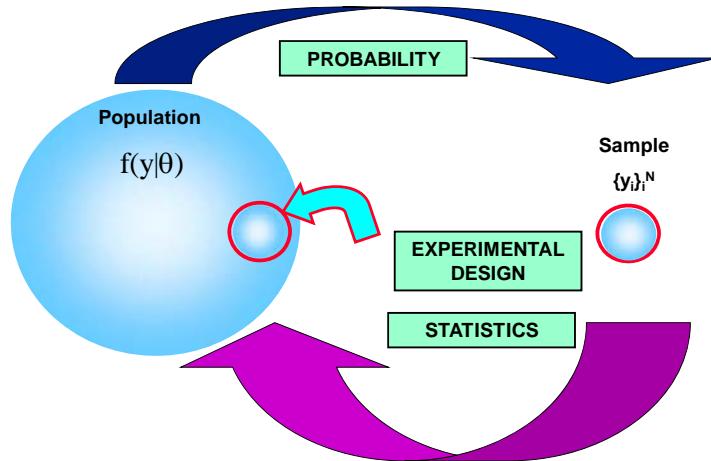
Now, suppose that this particular IVF treatment resulted in fraternal twins, i.e.  $x = 2$  (two individual embryos developed and were successfully delivered). This value is considered to be a single sample from a population that can be understood in one of two ways: (i) from a so-called “frequentist” perspective, the population in this case is the set of all actual IVF treatment outcomes one would obtain were it possible to repeat this three-embryo transfer treatment an infinite number of times on the same patient; (ii) The population could also be conceived equivalently as the collection of the outcomes of the same three-embryo IVF treatment carried out on an infinite number of *identically* characterized patients. In this regard, the 10-observation sample  $S_2$  would result from “sampling” nine more of such patients treated identically, whose pregnancy outcomes are 1, 1, 0, 3, 2, 2, 1, 2, 1, in addition to the already noted outcome  $x = 2$  from the first patient.

With this more practical problem, as with the coin-toss experiments, the pdf is known, but the parameter  $p$  is unknown; data is available, but in the form of finite-sized samples, whereas the full ensemble characterization we seek is of the entire population. We are left with no other choice but to use the samples, even though finite in size, to characterize the population. In these illustrations, this amounts to determining, from the sample data, a reasonable estimate of the true but unknown value of the parameter,  $p$ , for the specific problem at hand.

Observe therefore that in solving practical problems, the *population* — the full observational manifestation of the random variable,  $X$  — is that ideal, conceptual entity one wishes to characterize. The objective of random phenomena analysis is to characterize it completely with the pdf  $f(x|\boldsymbol{\theta})$ , where  $\boldsymbol{\theta}$  represents the parameters characteristic of the specific population in question. However, because it is impossible to realize the population in its entirety via experimentation, one must therefore settle for characterizing it by drawing *statistical inference* from a finite *sample* subset. Clearly, the success of such an endeavor depends on the sample being “representative” of the population. Statistics therefore involves not only systematic *analysis* of (necessarily finite) data, but also systematic data *collection* in such a way that the sample truly reflects the population, thereby ensuring that the sought-after information will be contained in the sample.

### 12.1.3 Probability, Statistics and Design of Experiments

We are now in a position to connect all these concepts into a coherent whole as illustrated in Fig 12.1. Associated with every random variable,  $X$ , is a pdf,  $f(x|\boldsymbol{\theta})$ , consisting of a functional form determined by the underlying random phenomenon and a set of characteristic parameters,  $\boldsymbol{\theta}$ . In general, given the complete  $f(x|\boldsymbol{\theta})$ , probability theory provides the tools for computing the probabilities of observing various occurrences  $\{x_i\}_{i=1}^n$ . For any *specific*



**FIGURE 12.1:** Relating the tools of Probability, Statistics and Design of Experiments to the concepts of Population and Sample

problem, however, only a finite set of observations  $\{x_i\}_{i=1}^n$  is available; the specific characterizing parameter vector,  $\theta$ , is unknown and must first be determined before probabilistic analysis can be performed. By considering the available finite set of observations as a sample from the (idealized) *complete* observational characterization of the random variable  $X$ , i.e. the population, the tools of Statistics make it possible to characterize the population fully (determine the functional form and estimate the unknown parameter in the pdf  $f(x|\theta)$ ) from information contained in the sample.

Finally, because the population is to be characterized on the basis of the finite sample, Design of Experiment provides the tools for ensuring that the sample is indeed representative of the population so that the information required to characterize the population adequately is contained in the sample.

#### 12.1.4 Statistical Analysis

It is customary to partition Statistics into two categories

1. Analyzing data (sometimes referred to as *Descriptive Statistics*), and,
2. Drawing inference from data (*Inductive* or *Inferential Statistics*),

although it is clear from above that basic to these partitions in a fundamental manner is a third category:

3. Generating data (Statistical Design of Experiments).

### Descriptive Statistics

In descriptive statistics, the primary concern is the presentation, organization and summarization of sample data, with emphasis on the extraction of information contained in a specific sample. Because the issue of generalization from sample to population does not arise, no explicit consideration is given to whether or not the sample is representative of the population. There are two primary aspects:

1. *Graphical*: The use of various graphical means of organizing, presenting and summarizing the data;
2. *Numerical*: The use of numerical measures, and characteristic values to summarize data.

Such an approach to data analysis is often referred to as “Exploratory Data Analysis (EDA)”<sup>1</sup>.

### Inductive Statistics

Also known variously as “Statistical Inference” or “Inferential Statistics,” it is primarily concerned with drawing inference about the population from sample data. As such, the focus is on generalization from the current data set (sample) to the larger population. The two primary aspects of inductive statistics are:

1. *Estimation*: Determining from sample data appropriate values for one or more unknown parameters of the assumed population description, the pdf;
2. *Hypothesis Testing*: Testing the validity of the assumed population model and the reliability of the parameter estimates.

### Statistical Design of Experiments

Whether for descriptive or inductive purposes, how sample data sets are obtained will affect the information they contain, and hence will influence our ability to generalize such information to the population *with confidence*. Design of Experiments (DoE) involves methods for efficient experimental procedures, data collection and analysis so that the sample may be considered as adequate. The assumptions underlying statistical analysis are thus rendered reasonable so that the extracted information may be as broadly applicable as possible.

<sup>1</sup>J. Tukey, (1977) *Exploratory Data Analysis*, Addison-Wesley.

The rest of this chapter is devoted to an overview treatment of “Descriptive Statistics;” a detailed discussion of “Inductive Statistics” follows in Chapters 13-18; and the central concepts and methodologies of Design of Experiments are discussed in Chapter 19. Our treatment of Statistics concludes with case studies in Chapter 20.

---

## 12.2 Variable and Data Types

Before embarking on a study of statistical techniques, it is important first to have a working knowledge of variable and data types.

Variables (and the data associated with them) are generally of two types:

1. *Quantitative Variables*: are numerical in nature, generating observations in the form of *numbers* that can be real-valued, (in which case the variable as well as the data are said to be *continuous*), or integers, in which case they are said to be *discrete*. The yield variables and data of Chapter 1 are continuous; the inclusions variable and data are discrete.
2. *Qualitative Variables*: are non-numeric and hence non-quantitative; the associated data are non-numeric observations. These variables tend to be descriptive, employing text, words, or symbols to convey a sense of “general meaning” rather than a precise measure of a quantity. Qualitative variables are sometimes referred to as “Categorical variables” because they typically describe categories: for example, if the variable is the *type* of an object under study, then the data, {Catalyst A, Catalyst B, Catalyst C} or {Fertilizer, No-Fertilizer} are qualitative; if the variable is the color of an object, then {Red, Green, Blue, Yellow} is a possible data set. The variable “opinion regarding the quality of a product,” may be associated with such entities as {Poor, Fair, Good, Better, Best}.

To be sure, most scientific investigations involve quantitative variables and data almost exclusively; but by no means should this be construed as implying that qualitative data are therefore unimportant in Science and Engineering. Apart from studies involving qualitative variables directly (“Drug A” versus a “Placebo”; Process A, or Process B; Machine 1, Machine 2 and Machine 3), naturally quantitative data are sometimes deliberately represented in qualitative form when this is required by the study. For example, even though one can quantify the “Ultimate Tensile Strength” (UTS) of a class of polymer resins with a precise numerical value (measured in MPa), it is actually quite common for a customer who is purchasing the product for further processing to classify the product simply as “Acceptable” if  $8 \text{ MPa} < UTS < 10 \text{ MPa}$  and “Unacceptable,” otherwise. The quantitative UTS data has been converted to qualitative data consisting of two categories. Similarly, while a quantitative

observation of the scores on an examination of 3 college roommates can be recorded as the set of numbers {98%, 87%, 63%}, a qualitative record of the same information may be presented as {High, Medium and Low} for each of the scores respectively, if the study is concerned more with categorizing scores than with quantifying them numerically.

The converse is also true: intrinsically qualitative data can sometimes be “represented” with numerical values. For example, opinion surveys often involve presenting a statement (such as: “*Chemical engineering professors have no sense of humor*”) to various individuals, each of whom is then asked to state his/her opinion by selecting from the following options: {1=Strongly Agree; 2=Agree; 3=Don’t Know; 4=Disagree; 5=Strongly Disagree}. The information gathered is intrinsically qualitative (a record of subjective opinion), and the assignment of the integers 1–5 is somewhat arbitrary; the same objective could have been achieved by assigning the numbers –2, –1, 0, 1, 2, or any other set of 5 distinct numbers. This brings us to another set of terms used to classify data.

1. *Nominal data*: have no order; they merely provide *names* or some other such identifying labels to various categories. For example, the set of weather conditions {Drizzle, Hail, Rain, Sleet, Snow, Thunderstorm} is *nominal*, as is the following set of manufactured personal care products {Soaps, Deodorant, Toothpaste, Shampoo}. There is no order implied or intended in such listings.
2. *Ordinal data*: have order, but the interval between each entry is meaningless. For example, the familiar classification {Small, Medium, Large} is *ordinal*; it is understood to indicate an increase in magnitude from the first to the last, but the difference between “Small” and “Medium” or that between “Medium” and “Large” is unspecified, neither is there any intention to indicate that one difference is of the same magnitude as the other. Similarly, the set of opinion poll options given above, {1=Strongly Agree; 2= Agree; 3=Don’t Know; 4=Disagree; 5=Strongly Disagree} is ordinal, indicating a generally declining level of “agreement” (equivalently, an increasing level of “disagreement”) of the subject with the validity of the statement in question. Note that the assigned numbers are meant to indicate no more than simple order: there is no intention that the “distance” between one entry and the other means anything.

Finally, it must be noted that while such nominal/ordinal classifications are valid in general, unusual exceptions sometimes arise in special fields of study. For example, the set of colors {Red, Orange, Yellow, Green, Blue, Indigo, Violet} is, under normal circumstances, entirely nominal; in Physics (Optics, specifically), however, this set is in fact ordinal, indicating the order of the components of visible light. The context in which such categorical data is presented usually makes the appropriate classification clear.

**TABLE 12.1:** Number and Type of injuries incurred by welders in the USA from 1980-1989

| Type of Injury | Total Number | Percent (%) |
|----------------|--------------|-------------|
| Eye            | 242          | 40.3        |
| Hand           | 130          | 21.7        |
| Arm            | 64           | 10.7        |
| Back           | 117          | 19.5        |
| Other          | 47           | 7.8         |

## 12.3 Graphical Methods of Descriptive Statistics

### 12.3.1 Bar Charts and Pie Charts

Consider the data shown in Table 12.1, a compilation of the number of recordable injuries incurred by members of the Welders Association of America, during the decade from 1980-1989, categorized by type. The variable “Type of Injury” is categorical, and hence qualitative; in addition, it is nominal because there is no particular order to the list. The other two variables, “Total Number” and “Percent” are quantitative: “Total number” is discrete since it is a count; “Percent” is continuous by virtue of being a ratio of integers.

This data set can be represented graphically several different ways. A *bar chart* (or *bar graph*) is a means by which the numbers associated with each category are represented by rectangular bars whose heights are proportional to the respective numbers. Such a chart is shown in Fig 12.2 for the total number of injuries associated with each injury type. This is a vertical bar chart because the bars are oriented vertically, as opposed to horizontal charts where the bars are oriented horizontally.

Note that because the “Type” variable is ordinal, there is no particular order in which the data should be represented. Under these circumstances, to avoid the somewhat haphazard impression one gets from Fig 12.2, it is often recommended to order the bars in some meaningful manner, typically by ranking the plotted values in increasing or decreasing order. When the values in the default Fig 12.2 plot are ranked in decreasing order, one obtains Fig 12.3. One advantage of such a rearrangement is that should there arise an interest in instituting a program to prevent injuries, and one can only focus on a few injury types at a time, the figure provides an easy visual representation that can be used objectively to convey the logic behind the prioritization. For example, the figure indicates that Eye, Hand, and Back injuries make up most of the injuries, with the majority contributed by Eye injuries so that if there are resources for tackling only one injury, the logical choice is obvious.

This sort of analysis — involving the relative contribution of various cat-

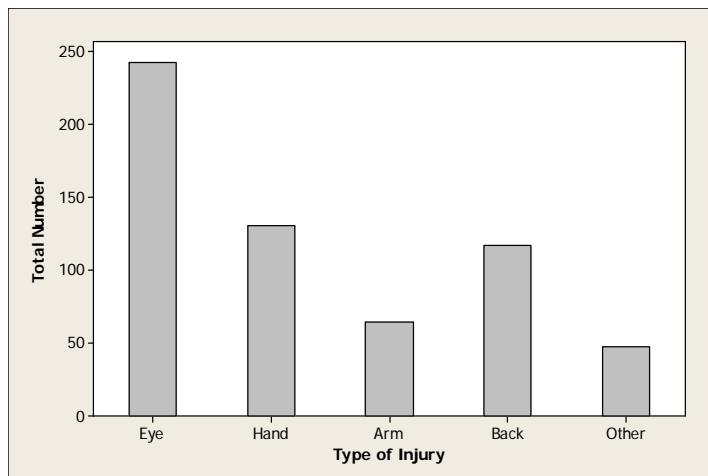


FIGURE 12.2: Bar chart of welding injuries from Table 12.1

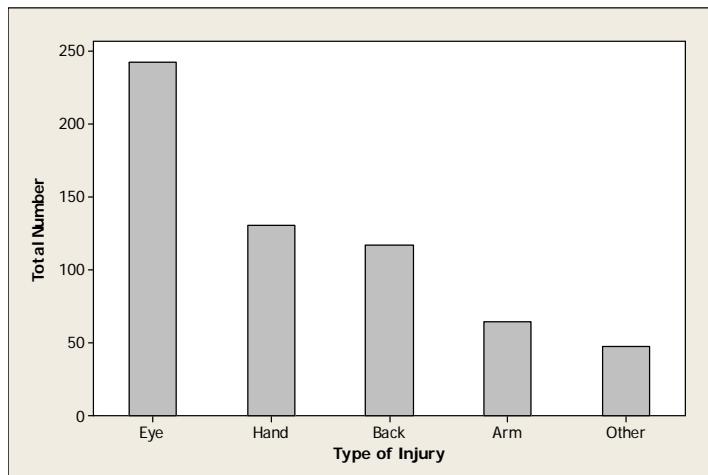


FIGURE 12.3: Bar chart of welding injuries arranged in decreasing order of number of injuries

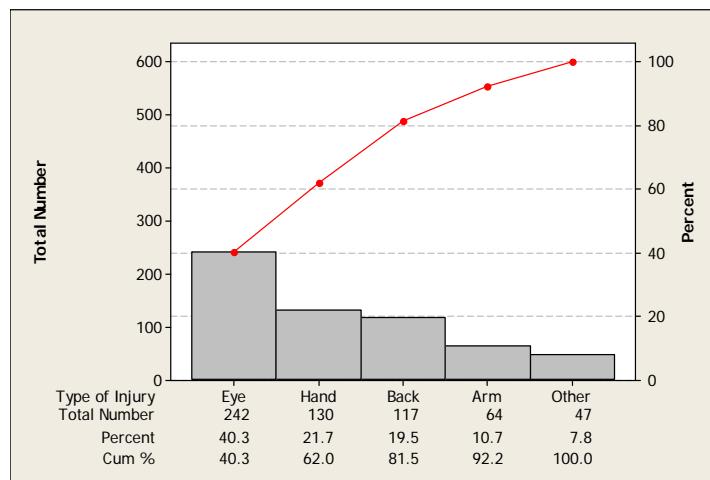


FIGURE 12.4: Pareto chart of welding injuries

egories (or factors) to a variable of interest — is known as a *Pareto Analysis*, named after Vilfredo Federico Pareto (1848–1923), the Italian economist and sociologist best known for his discovery that 80% of Italy's wealth was owned by 20% of the population. This observation has since been generalized to other phenomena in which 80% of some category of “effects” is due to 20% of the causes. It has become particularly useful in industrial quality assurance where the primary concern is to determine the “vital few” causes of poor quality upon which to concentrate quality improvement programs.

When a rank-ordered bar chart of the sort in Fig 12.3 is accompanied by a plot of the cumulative total contribution from each category, the result is known as a Pareto chart. Such a chart for the welding injury data is shown in Fig 12.4.

An alternative to using bars to represent frequencies of categorical data is the “Pie chart,” a graphical technique that is very popular in journalism — print and electronic — and in business. With a pie chart, the categories are represented as wedges of a “pie” or sectors of a circle, whose areas (or equivalently, arc lengths or central angles) are proportional to the values (or relative frequencies) associated with each category. For example, the pie chart for the welding injury data is shown in Fig 12.5.

Several variations of this basic chart form exist (“exploded” pie chart; “perspective” pie charts, etc) but we will not discuss any of these here. The pie chart is used less often in engineering and science primarily because comparing areas visually is more difficult than comparing lengths. (The items in Table 12.1 are easier to compare visually in Fig 12.2 than in Fig 12.5.) It is also more difficult to compare information in different pie charts as opposed to the same information presented in different bar charts. The pie chart is best used

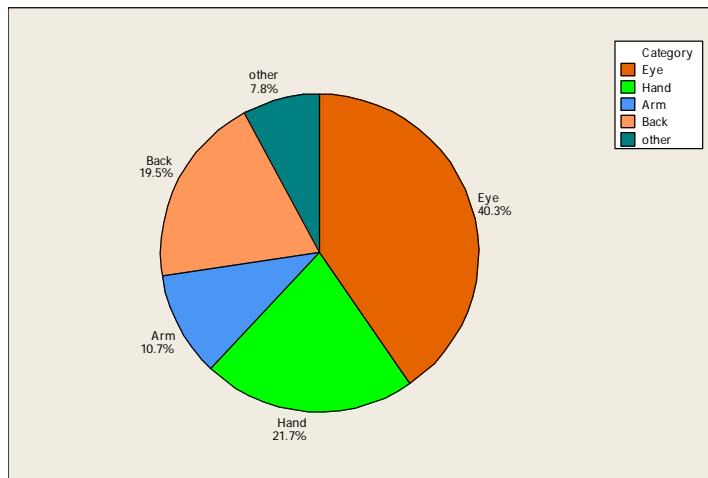


FIGURE 12.5: Pie chart of welding injuries

**TABLE 12.2:** Frozen Ready meals in France, in 2002

| Type of Food          | Percent |
|-----------------------|---------|
| French Regional Meals | 29.5    |
| Cooked Fish           | 25.0    |
| Pasta Based Meals     | 16.6    |
| European Meals        | 9.4     |
| Side Dishes           | 6.8     |
| Cooked Seafood        | 6.3     |
| Exotic                | 4.7     |
| Cooked Snails         | 1.7     |

when one is primarily concerned with comparing a particular category with the entire group (for example, that eye injuries contribute the largest to the collection of welding injuries is very clearly depicted in Fig 12.5).

The following example uses actual data to illustrate the strengths and weaknesses of the bar and pie charts.

**Example 12.2 COMPARING BAR CHARTS AND PIE CHARTS: FROZEN READY MEALS SOLD IN FRANCE IN 2002**

The data in Table 12.2, from the *Global Foods Almanac*, October, 2002, summarizes 2002 sales information about frozen ready meals in France. Generate a bar chart and a pie chart of the data and briefly compare the charts.

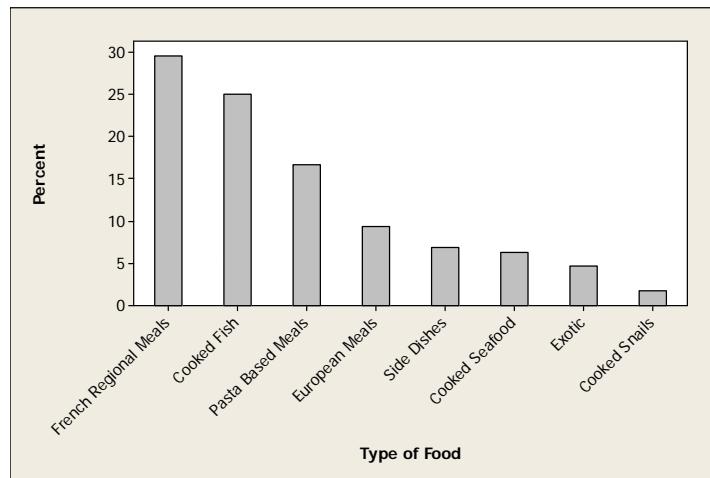


FIGURE 12.6: Bar Chart of frozen ready meals sold in France in 2002

**Solution:**

The bar chart is shown in Fig 12.6, the pie chart in Fig 12.7. Keeping in mind that they are primarily used for visual communication of numerical facts, here are the most salient aspects of these charts: (i) It is *not* very easy to see from the pie chart wedges (without the added numbers) which category, *French Regional Meals*, or *Cooked Fish*, accounts for the larger proportion of the overall sales, since both wedges look about the same in size; with the bar chart, there is no question, even if the bars in question were not situated side-by-side. (ii) However, if one is familiar with reading pie charts, it is far easier to see that the “Cooked Fish” category accounts for *precisely* a quarter of the total sales (the right angle subtending the “Cooked Fish” wedge is the key visual cue); this fact is not at all obvious from the bar chart which is much less effective at conveying “relative-to-the-whole” information. (iii) *French regional meals* sold approximately 20 times more than *Cooked snails*; even with the attached numbers, this fact is much easier to appreciate in the bar chart than in the pie chart.

Thus, observe that while the pie chart excels at conveying “relative-to-the-whole” information (especially if the relative proportions in question are 25%, 50%, or 75% — entities whose angular representations are easily recognizable by the unaided eye), the bar chart is weak in this regard. Conversely, the bar chart conveys “relative-to-one-another” information far better than the pie chart.

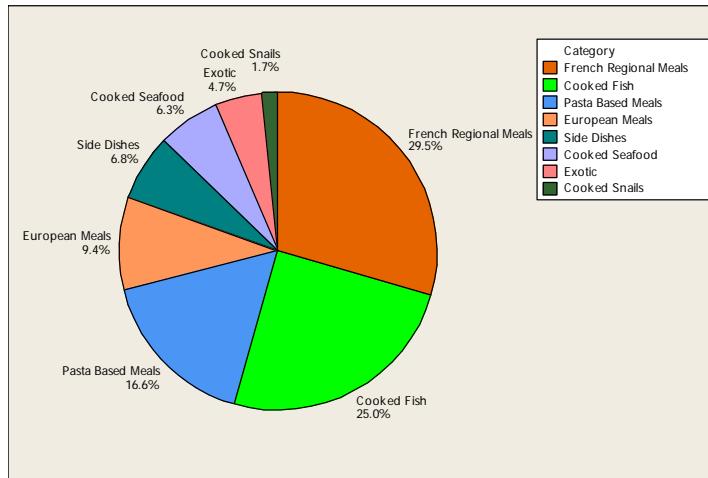


FIGURE 12.7: Pie Chart of frozen ready meals sold in France in 2002

### 12.3.2 Frequency Distributions

As previewed in Chapter 1, even quantitative data can be quite uninformative if presented in raw form, as numbers in a table. One of the first steps in making sense of the information contained in raw quantitative data is to rearrange the data, dividing them into smaller groups, or classes (also known as “bins”) and attaching to each group a number representing how many of the raw data belong to that group (i.e. how frequently members of that group occur in the raw data set). The result is a “frequency distribution” representation of the data. When the frequency  $\phi_i$  associated with each group  $i$  is normalized by the total number of data points,  $n$ , in the sample set, we obtain the “relative frequency,”  $f_i$ , for each group. For example, a specific reorganization of the yield data,  $Y_A$ , presented in Chapter 1 gives rise to the frequency distribution shown in Table 1.3 of Chapter 1 and reproduced here (in Table 12.3) for easy reference. Compared to the raw data in Table 1.1, this is a more compact and more informative representation of the original data. Of course, such compactness is achieved at the expense of some details, but this loss is more than compensated for by a certain enhanced clarity with which the true character of the random variation begins to emerge from this frequency distribution. For example, the frequency distribution shows clearly how much of the data clusters around the group [74.51-75.50], an important characteristic that is not readily evident from the raw data table.

A plot of this frequency distribution using vertical bars whose heights are proportional to the frequencies (or, equivalently, the relative frequencies) is known as a *histogram*, with the one corresponding to the  $Y_A$  frequency distribution shown in Fig 1.1 and repeated here (in Fig 12.8) for ease of reference.

**TABLE 12.3:** Group classification  
and frequencies for  $Y_A$  data

| $Y_A$ group | Frequency<br>$\phi_i$ | Relative<br>Frequency<br>$f_i$ |
|-------------|-----------------------|--------------------------------|
| 71.51-72.50 | 1                     | 0.02                           |
| 72.51-73.50 | 2                     | 0.04                           |
| 73.51-74.50 | 9                     | 0.18                           |
| 74.51-75.50 | 17                    | 0.34                           |
| 75.51-76.50 | 7                     | 0.14                           |
| 76.51-77.50 | 8                     | 0.16                           |
| 77.51-78.50 | 5                     | 0.10                           |
| 78.51-79.50 | 1                     | 0.02                           |
| TOTAL       | 50                    | 1.00                           |

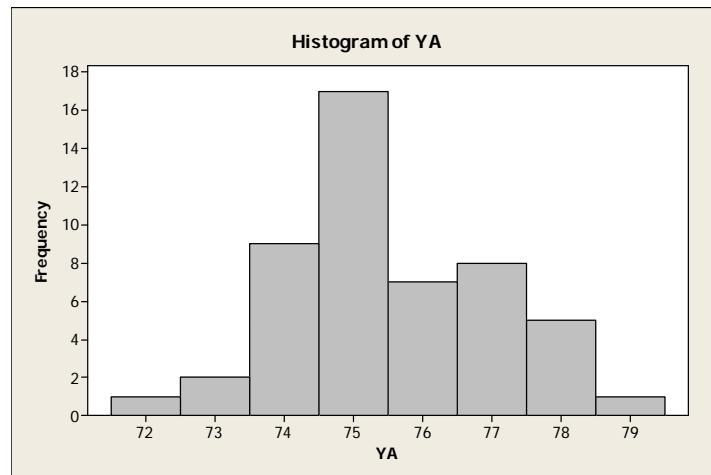


FIGURE 12.8: Histogram for  $Y_A$  data of Chapter 1

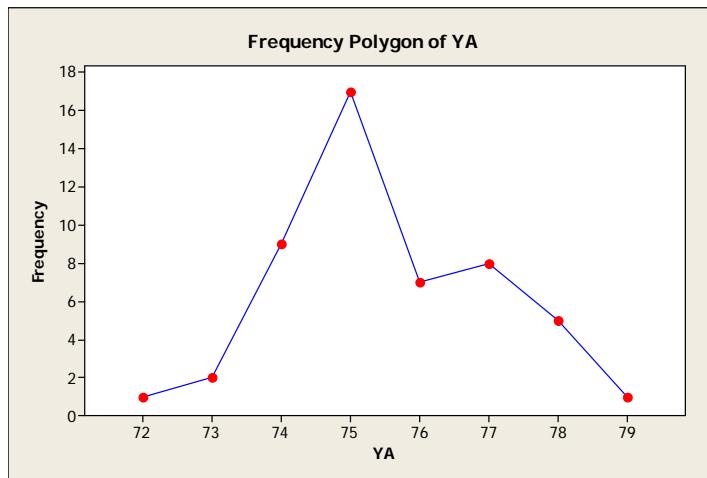
Although by far the most popular, the histogram is not the only graphical means of representing frequency distributions. If, instead of using adjoining bars to represent group frequencies, we employed a cartesian plot in which each group is represented by its center value on the  $x$ -axis and the corresponding frequency plotted on the  $y$ -axis, with the points connected by straight lines, the result is known as a “frequency polygon” as shown in Fig 12.9 for the  $Y_A$  data. This is only a slightly different rendering of the information contained in the histogram. Fig 12.10 shows the corresponding frequency polygon for the  $Y_B$  data of Chapter 1 (whose histogram is shown in Fig 1.4).

As alluded to in Chapter 1, such graphical representations of the data provide an empirical — and partial, as opposed to complete — approximation to the true underlying distribution; but they show features not immediately apparent from the raw data. Because data sets are “incomplete samples,” the corresponding frequency distributions show irregularities; but, as the sample size  $n \rightarrow \infty$ , these irregularities gradually diminish, so that these empirical distributions ultimately approach the complete population distribution of the random variable in question. These facts inform the “frequentist” approach to statistics and data analysis.

Some final points to note about frequency distributions: It should be clear that to be meaningful, histograms must be based on an ample amount of data; only then will there be a sufficient number of groups, with enough members per group, to display the data distribution meaningfully. As such, whenever possible, one should avoid employing histograms to display data sets containing fewer than 15-20 data points.

It should also be clear that the choice of “bin size” will affect the general appearance of the resulting histogram. Bins that are too wide generate fewer groups and the resulting histograms cannot adequately reveal the true distributional characteristics of the data. As an extreme example, a bin size covering the entire range of a data set containing a total of  $n$  data points will produce a histogram consisting of a single vertical bar of height  $n$ . Of course, this is totally uninformative — at least no more informative than the raw data table — because the entire data set will remain confined to this one group. On the other extreme, if the bin size is so small that, with the exception of exactly identical data entries, each data point fits into a group all by itself, the result is an equally uninformative histogram — uninformative for a complementary reason: this time, the entire data set is stretched out horizontally into a collection of  $n$  bars, all of the same unit height. Somewhere between these two obviously untenable extremes lies an acceptable bin size, but there are no hard-and-fast rules for choosing it. The rule-of-thumb is that any choice resulting in  $\sim 8 - 10$  groups is considered as acceptable.

In practice, transforming raw data sets into frequency distributions and the accompanying graphical representations is almost always carried out by computer programs such as MINITAB, Matlab, SAS, etc. And these software packages are preprogrammed with algorithms that automatically choose reasonable bin sizes for each data set. The traditional recommendation for the

FIGURE 12.9: Frequency Polygon of  $Y_A$  data of Chapter 1

number of intervals,  $k$ , to use in representing a data sample of size  $n$  is the following, from Sturges, 1926<sup>2</sup>,

$$k = 1 + 3.3 \log_{10} n \quad (12.6)$$

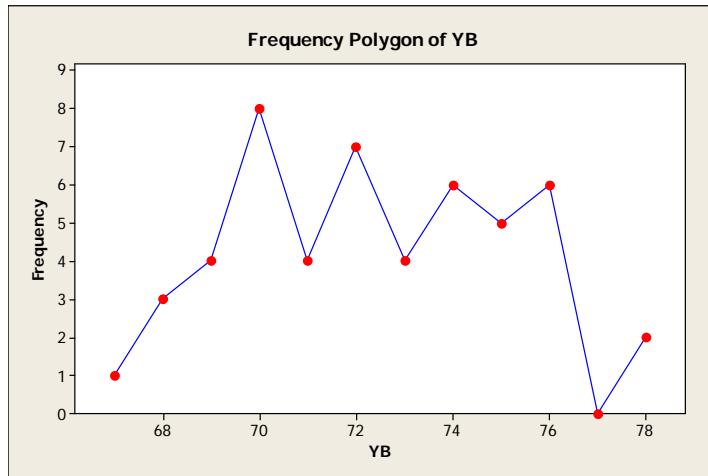
Thus, for instance, for the yield data, with  $n = 50$ , the recommendation will be 7 intervals. The histogram in Fig 12.8, generated automatically with MINITAB, uses 8 intervals.

### 12.3.3 Box Plots

Also known as a “box-and-whisker” plot, the box plot was proposed in 1977 by the American statistician John Tukey (1915-2000) as a means of presenting, in a single compact plot, the following five key characteristics of a data set:

1. The minimum (smallest valued observation);
2. The lower (or first) quartile,  $Q_1$ , (25% of the data values are less than or equal to this value);
3. The median, or middle value, (50% of the data values are less than or equal to this value, and 50% are greater than or equal to it);
4. The upper (or third) quartile,  $Q_3$ , (75% of the data values are less than or equal to this value; equivalently, 25% of the data are greater than or equal to this value);

<sup>2</sup>Sturges, H.A., (1926). “The choice of a class interval,” J. Am. Stat. Assoc., 21, 65-66.

FIGURE 12.10: Frequency Polygon of  $Y_B$  data of Chapter 1

5. The maximum (largest valued observation).

These items are also known as the “five-number summary” of a data set. What gives the plot its name is how these 5 characterizing quantities are depicted in the form of a rectangular box and two “whiskers” extending from the two ends, as described below:

When oriented vertically, the bottom of the box represents the first quartile,  $Q_1$ , the top of the box is the third quartile,  $Q_3$ , while a middle line inside the box represents the median. This vertical box therefore encompasses the lower and upper quartiles of the data set so that its length is the interquartile range,  $IQR = Q_3 - Q_1$ .

How the data minimum and maximum are depicted in this graphical representation is another of its attractive features. Based on a supposition in normally distributed data sets, extreme values (minimum and maximum) hardly fall outside a region that is 1.5 times the interquartile range from the lower or upper quartile, box plots employ an upper and a lower limit defined as follows:

$$UL = Q_3 + 1.5(Q_3 - Q_1) \quad (12.7)$$

$$LL = Q_1 - 1.5(Q_3 - Q_1) \quad (12.8)$$

With this definition, the lower whisker is drawn as a line extending from the bottom of the box (i.e from  $Q_1$ ) to the smallest data value so long as it falls within the lower limit. In this case, the end of the bottom whisker is therefore the data minimum. Any data value that falls outside the lower limit is flagged as a potential “outlier” — in this case, an unusually small observation — and represented with an asterisk. The upper whisker is drawn similarly: from the top of the box,  $Q_3$ , to the largest data value within the upper limit. All data

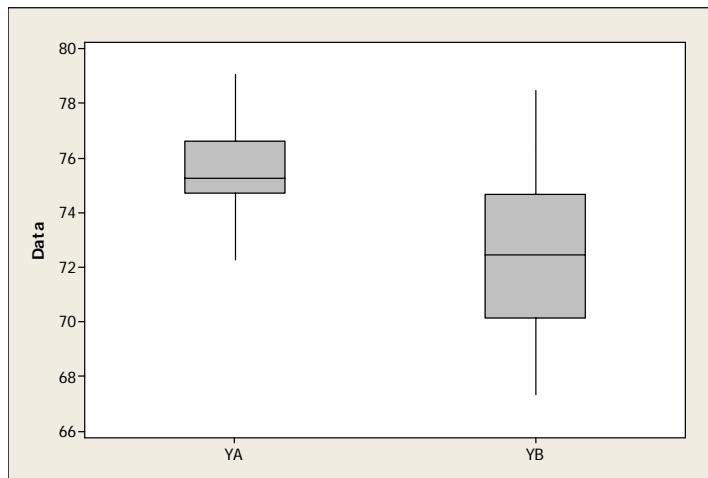


FIGURE 12.11: Boxplot of the chemical process yield data  $Y_A$ ,  $Y_B$  of Chapter 1

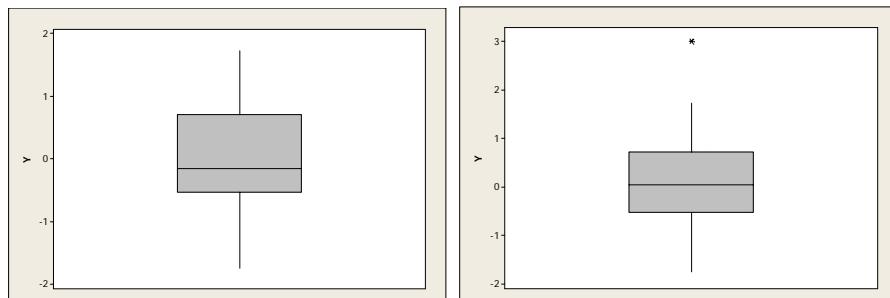
values exceeding the upper limit are also flagged as outliers — unusually large observations in this case — and also represented with an asterisk.

Figure 12.11 shows box plots for the chemical process yield data sets  $Y_A$  and  $Y_B$  introduced in Chapter 1. Observe how box plots are particularly adept at displaying data sets visually without overwhelming detail. In one compact plot, they show clearly the data range, symmetry, central location, and concentration around the center. They are also good for quick visual comparisons of data sets. Thus, for example, the details contained in the yield data sets are compressed succinctly in these plots, showing, among other things, that both data sets are essentially symmetric about their respective centers which is higher for  $Y_A$  than for  $Y_B$ ; the minimum for  $Y_B$  is substantially lower than that for  $Y_A$ ; in fact, the “central 50%” of the  $Y_A$  data, the rectangular box on the left, is much more compact and entirely higher than the corresponding “central 50%” for  $Y_B$ . Also the overall comparative compactness of the  $Y_A$  plot indicates visually how this data set is less variable overall than  $Y_B$ . Of course, all of this information is available from the histogram, but while a relatively large number of observations are required in order for a histogram to be meaningful, a data set consisting of as little as 5 observations can be represented meaningfully in a box plot. Thus, while box plots are useful for all quantitative data, they are especially good for small data sets.

Figure 12.12 shows on the left a box plot of 30 observations from a random variable  $Y \sim N(0, 1)$ ; shown on the right is the same data set with a single addition of the number 3.0 as the 31<sup>st</sup> observation. Note that this value is flagged as unusually high for this data set.

### Example 12.3 RAISIN-DISPENSING MACHINES

In a study to determine the performance of 5 different processing ma-



**FIGURE 12.12:** Boxplot of random  $N(0,1)$  data: original set, and with added “outlier”

**TABLE 12.4:** Number of raisins dispensed into trial-sized “Raisin Bran” cereal boxes by five different machines

| Machine 1 | Machine 2 | Machine 3 | Machine 4 | Machine 5 |
|-----------|-----------|-----------|-----------|-----------|
| 27        | 17        | 13        | 7         | 15        |
| 21        | 12        | 7         | 4         | 19        |
| 24        | 14        | 11        | 7         | 19        |
| 15        | 7         | 9         | 7         | 24        |
| 33        | 14        | 12        | 12        | 10        |
| 23        | 16        | 18        | 18        | 20        |

chines used to add raisins to a trial-size “Raisin Bran” cereal boxes, 6 sample boxes are taken at random from each machine’s production line and the number of raisins in each box counted. The result is displayed in Table 12.4.

Because the number of samples for each machine is so few, individual histograms of each machine’s data will not be meaningful. Instead, obtain five individual box plots for this data set. What do these plots suggest about the equality of how these machines dispense raisins into each box?

**Solution:**

The box plots are shown in Fig 12.13, from which it appears as if there are some noticeable differences in how this group of machines dispense raisins. In particular, the plots seem to suggest that Machine 1 (and possibly Machine 5) may be dispensing more raisins than the others, while the other three machines appear to be similar in the way they dispense raisins.

These somewhat “informal” and “descriptive” statements can be made more rigorous and quantitative using techniques for drawing precise statistical inferences about the significance (or otherwise) of any observed differences. Such techniques are discussed in greater detail in Chapter 15.

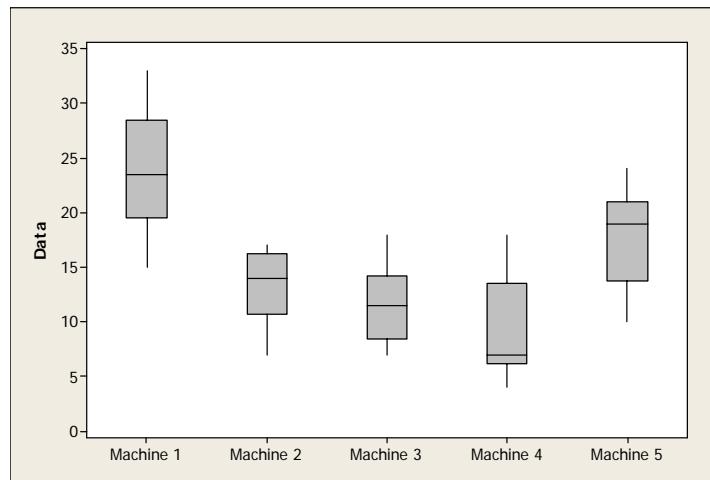


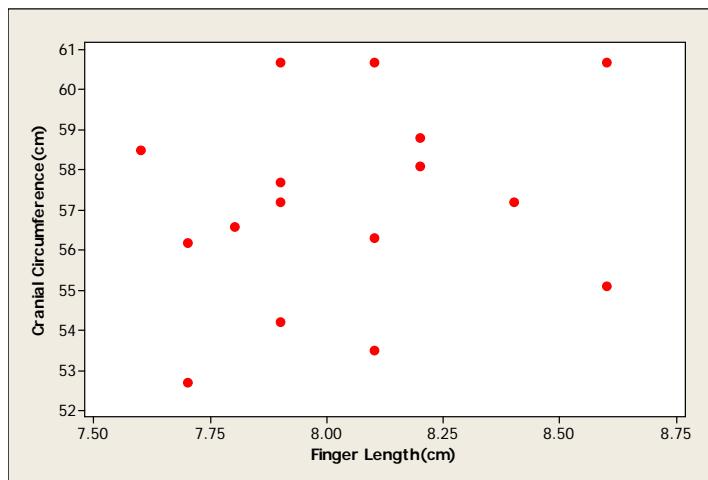
FIGURE 12.13: Box plot of raisins dispensed by five different machines

#### 12.3.4 Scatter Plots

When the values of one variable are plotted on the  $y$ -axis versus the values of another variable on the  $x$ -axis, the result is known as a *scatter plot*. The plot is so-called because, unless the one variable is perfectly correlated with the other, the data appear “scattered” in the plot. Such plots provide visual clues as to whether or not there truly is a relationship between the variables, and if there is one, what sort of relationship — strong or weak, linear or nonlinear, etc. Although not necessarily always the case, the variable plotted on the  $y$ -axis is usually the one that may potentially be “responding” to the *other* variable, which will be plotted on  $x$ -axis. It is also possible that a causal relationship may not exist between the plotted variables, or if there is one, it may not always be clear which variable is responding and which is causing the response. Because these plots are truly just “exploratory,” care should be taken not to over-interpret them; it is especially important not to jump to conclusions that any observed apparent relationship implies causality.

A series of scatter plots are shown below, beginning with Fig 16.6, a plot of the cranial circumference and corresponding finger length for various individuals. Believe it or not, there once was a time when people speculated that these two variables correlate. This plot shows that, at least for the individuals involved in the particular study generating the data, there does not appear to be any clear relationship between these two variables. Even if the plot had indicated a strong relationship between the variables, observe that in this case, none of these two variables can be rightly considered as “dependent” on the other; thus, the choice of which variable to plot on which axis is purely arbitrary.

Next, consider the data shown in Table 12.5, city and highway gasoline



**FIGURE 12.14:** Scatter plot of cranial circumference versus finger length: The plot shows no real relationship between these variables

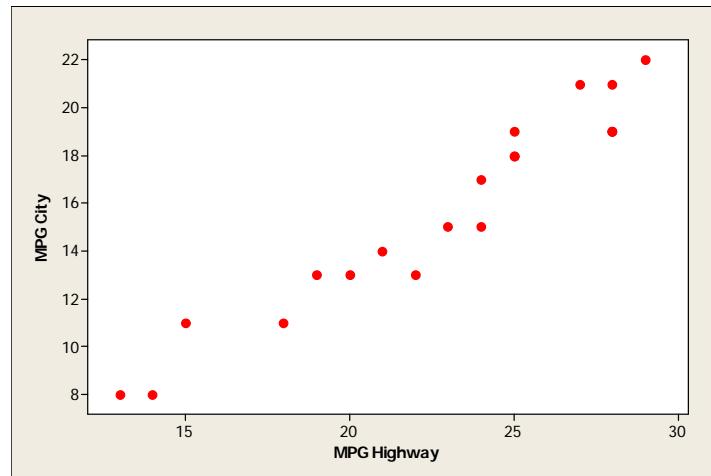
mileage ratings, in miles per gallon (mpg), for 20 types of two-seater automobiles, complete with engine characteristics, capacity (in liters) and number of cylinders. First, a plot of city gas mileage versus highway gas mileage, shown in Fig 12.15, indicates a very strong, positive linear relationship between these two variables. However, even though related, it is clear that this relationship is not causal in the sense that one cannot “independently and directly” manipulate say city gas mileage and as a direct consequence thereby *cause* highway gas mileage to change. Rather, both variables depend in common on other factors that can be independently and directly manipulated (e.g., engine capacity).

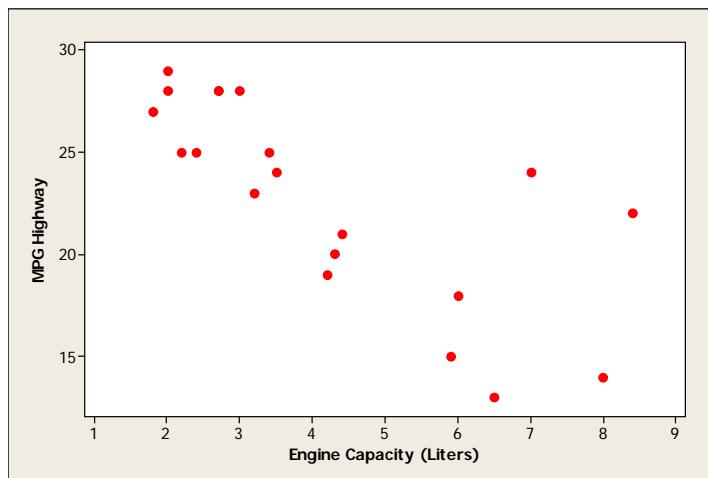
Fig 12.16 shows a plot of highway gas mileage against engine capacity, indicating an approximately linear and negative relationship. Observe that according to the thermodynamics of internal combustion engines, the physics of work done by applying force to move massive objects, and the fact that larger engines are normally required for bigger cars, it is entirely logical that smaller engines correlate with higher highway gas mileage. Fig 12.17 shows a corresponding plot of highway gas mileage versus the number of cylinders. This plot also indicates a similar negative, and somewhat linear relationship between the variables. (Because city and highway gas mileage values are so strongly correlated, similar plots of the city mileage data should show characteristics similar to the corresponding highway mileage plots. See Exercise 12.15)

Scatter plots are also very good at pointing out data that might appear “inconsistent” with others in the group. For example, in Fig 12.16, two data points for engine capacities 7 liters and 8.4 liters are associated with highway

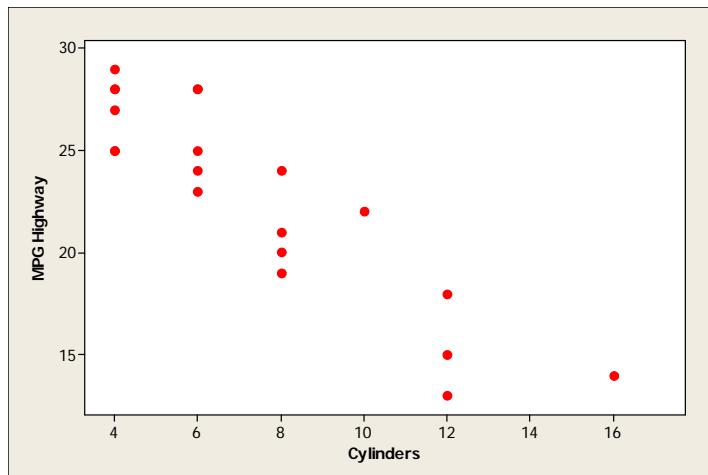
**TABLE 12.5:** Gasoline mileage ratings for a collection of two-seater automobiles

|    | Car Type and Model      | Eng Capacity (Liters) | # Cylinders | City mpg | Highway mpg |
|----|-------------------------|-----------------------|-------------|----------|-------------|
| 1  | Aston Marton V8 Vantage | 4.3                   | 8           | 13       | 20          |
| 2  | Audi R8                 | 4.2                   | 8           | 13       | 19          |
| 3  | Audi TT Roadster        | 2.0                   | 4           | 22       | 29          |
| 4  | BMW Z4 3.0i             | 3.0                   | 6           | 19       | 28          |
| 5  | BMW Z4 Roadster         | 3.2                   | 6           | 15       | 23          |
| 6  | Bugatti Veyron          | 8.0                   | 16          | 8        | 14          |
| 7  | Caddilac XLR            | 4.4                   | 8           | 14       | 21          |
| 8  | Chevrolet Corvette      | 7.0                   | 8           | 15       | 24          |
| 9  | Dodge Viper             | 8.4                   | 10          | 13       | 22          |
| 10 | Ferrari 599 GTB         | 5.9                   | 12          | 11       | 15          |
| 11 | Honda S2000             | 2.2                   | 4           | 18       | 25          |
| 12 | Lamborghini Murcielago  | 6.5                   | 12          | 8        | 13          |
| 13 | Lotus Elise/Exige       | 1.8                   | 4           | 21       | 27          |
| 14 | Mazda MX5               | 2.0                   | 4           | 21       | 28          |
| 15 | Mercedes Benz SL65 AMG  | 6.0                   | 12          | 11       | 18          |
| 16 | Nissan 350Z Roadster    | 3.5                   | 6           | 17       | 24          |
| 17 | Pontiac Solstice        | 2.4                   | 4           | 19       | 25          |
| 18 | Porsche Boxster-S       | 3.4                   | 6           | 18       | 25          |
| 19 | Porsche Cayman          | 2.7                   | 6           | 19       | 28          |
| 20 | Saturn SKY              | 2.0                   | 4           | 19       | 28          |

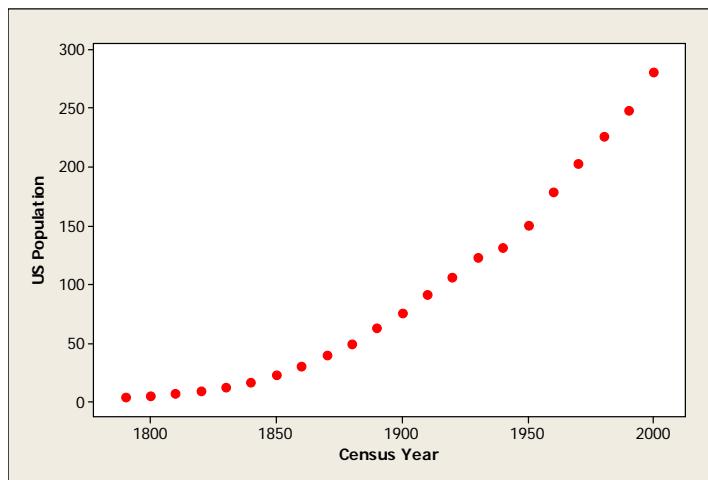
**FIGURE 12.15:** Scatter plot of city gas mileage versus highway gas mileage for various two-seater automobiles: The plot shows a strong positive linear relationship, but no causality is implied.



**FIGURE 12.16:** Scatter plot of highway gas mileage versus engine capacity for various two-seater automobiles: The plot shows a negative linear relationship. Note the two unusually high mileage values associated with engine capacities 7.0 and 8.4 liters identified as belonging to the Chevrolet Corvette and the Dodge Viper, respectively.



**FIGURE 12.17:** Scatter plot of highway gas mileage versus number of cylinders for various two-seater automobiles: The plot shows a negative linear relationship.



**FIGURE 12.18:** Scatter plot of US population every ten years since the 1790 census versus census year: The plot shows a strong non-linear trend, with very little scatter, indicative of the systematic, approximately exponential growth

gas mileage values of 24 and 22 miles per gallon, respectively — values that appear unusually high for such large engines, especially when compared to corresponding values for other automobiles with engines of similar volume. An inspection of the data table indicates that these values belong to the Chevrolet Corvette and the Dodge Viper models respectively — automobiles whose bodies are constructed from special fiberglass composites, resulting in vehicles that are generally lighter in weight than others in their class. The scatter plot correctly shows these data to be “inconsistent” with the rest, and we are able to provide a logical reason for the unusually high gas mileage: lighter cars, even those with large engines, generally get better gasoline mileage.

When the variable on the  $x$ -axis is time, the plot provides an indication of any trends that may exist in the  $y$  variable. Examples of such plots include monthly sales volume of a particular item; daily closing values of stocks on the stock exchange; monthly number of drunk driving arrests on a municipal road, etc. These are all plots that indicate time trends in the variables of interest.

Fig 12.18 shows a plot of the populations of the United States of America as determined by the decade-by-decade census from 1790 to 2000. This plot shows the sort of exponential growth trend that is typical of populations of growing organisms. We revisit this data set in Chapter 20.

## 12.4 Numerical Descriptions

Characterizing data sets by empirical frequency distributions, while useful for graphically condensing the information contained in the data into a relatively small number of groups, is not particularly useful for carrying out *quantitative* comparisons. Such comparisons require quantitative numerical descriptors of the data characteristics, typically measures of (i) central tendency (or data “location”); (ii) variability (or spread); (iii) skewness, and (iv) peakedness. It is not coincidental that these common numerical descriptors align perfectly with the characteristic moments of theoretical distributions discussed in Chapter 4. In statistical analysis, these numerical descriptors are computed from sample data as single numbers used to represent various aspects of the entire data set; they are therefore numerical approximations of the corresponding true but unknown population distribution parameters. Given sample data, all such numerical descriptors are, of course, routinely computed by various statistical analysis software packages; the following discussion simply provides some perspective on the most common of these descriptors. In particular, we demonstrate the sense in which they are to be considered as appropriate measures, and hence clarify the context in which they are best utilized.

### 12.4.1 Theoretical Measures of Central Tendency

Given a sample  $x_1, x_2, \dots, x_n$ , obtained as experimental observations of a random variable,  $X$ , we wish to consider how best to choose a number,  $c$ , to represent the “central location” of this random variable and its  $n$  observed realizations. Because  $X$  is a random variable, it seems entirely reasonable to choose this number such that the *expectation* of some function of the deviation term  $(X - c)$  is minimized. Let us therefore define

$$\phi_n = E [|X - c|^n] \quad (12.9)$$

and seek values of  $c$  that will minimize  $\phi_n$  for various values of  $n$ .

#### The Mean

When  $n = 2$ , the objective function to be minimized in Eq (12.9) becomes

$$\phi_2 = E [(X - c)^2] \quad (12.10)$$

which expands out to give:

$$\begin{aligned} \phi_2 &= E [X^2 - 2Xc + c^2] \\ &= E[X^2] - 2cE[X] + c^2 \end{aligned} \quad (12.11)$$

because  $c$  is a constant. Employing the standard tools of calculus — differentiating in Eq (12.11) with respect to  $c$ , setting the result to zero and solving for  $c$  — yields:

$$\frac{d\phi_2}{dc} = -2E[X] + 2c = 0 \quad (12.12)$$

giving the immediate (and not so surprising) result that

$$c = E[X] = \mu \quad (12.13)$$

Thus, the mean is the “best” single representative of the theoretical centroid of a random variable if we are concerned with minimizing mean squared deviation from all possible values of  $X$ .

### The Median

When  $n = 1$  in Eq (12.9), we wish to find  $c$  to minimize the mean absolute deviation between it and the possible values of the random variable,  $X$ . For the continuous random variable, by definition,

$$\phi_1 = \int_{-\infty}^{\infty} |x - c| f(x) dx \quad (12.14)$$

so that

$$\frac{\partial \phi_1}{\partial c} = \int_{-\infty}^{\infty} \frac{\partial}{\partial c} |x - c| f(x) dx = 0 \quad (12.15)$$

is the equation to be solved to find the desired  $c$ . Now, because  $|x - c|$  is discontinuous at the point  $x = c$ , the indicated differentiation must be carried out with caution; in particular, we note that

$$\frac{\partial}{\partial c} |x - c| = \begin{cases} -1; & x > c \\ 1; & x < c \end{cases} \quad (12.16)$$

As a result, Eq (12.15) becomes:

$$0 = \int_{-\infty}^c f(x) dx - \int_c^{\infty} f(x) dx \quad (12.17)$$

where the first term represents the integral over the region  $\{x : x < c\}$  and the second term is the integral over the remaining region  $\{x : x > c\}$ . This equation has the obvious solution:

$$\int_{-\infty}^c f(x) dx = \int_c^{\infty} f(x) dx, \quad (12.18)$$

which is either immediately recognized as the definition of  $c$  as the median of the pdf  $f(x)$ , or, by explicitly introducing the cumulative distribution function,  $F(x)$ , reduces to

$$F(c) = 1 - F(c), \quad (12.19)$$

which now yields

$$F(c) = 0.5; \Rightarrow c = x_m \quad (12.20)$$

where  $x_m$  is the median. It is left as an exercise for the reader (see Exercise 12.18) to establish the result for discrete  $X$ .

Thus, the median is the central representative of  $X$  that gives the smallest mean absolute deviation from all possible values of  $X$ .

### The Mode

It is shown in the Appendix at the end of this chapter that when  $n = \infty$ ,

$$c = x^* \quad (12.21)$$

where  $x^*$  is the mode of the pdf  $f(x)$ , minimizes the objective function in Eq (12.9). The mode is therefore the central representative of  $X$  that provides the smallest of the worst possible deviations from all possible values of  $X$ .

This discussion puts into perspective the three most popular measures of “central location” of a random variable — the mean, median, and mode — their individual theoretical properties and what makes each one a good measure. Theoretically, for all symmetric distributions, the mean, mode and median all coincide; they differ (sometimes significantly) for nonsymmetric distributions. The sample data equivalents of these population parameters are obtained as discussed next.

#### 12.4.2 Measures of Central Tendency: Sample Equivalents

##### Sample Mean

From a sample  $x_1, x_2, x_3, \dots, x_n$ , the sample mean, or the sample average,  $\bar{x}$ , is defined as:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (12.22)$$

In terms of the just-concluded theoretical considerations, this implies that of all possible candidate values,  $c$ , the sample average, is therefore that value which minimizes the mean squared error between the observed realizations of  $X$  and  $c$ . This quantity is sometimes referred to as the “arithmetic mean” to distinguish it from other means. For example, the *geometric mean*,  $\bar{x}_g$ , defined as

$$\bar{x}_g = \left[ \prod_{i=1}^n x_i \right]^{1/n}, \quad (12.23)$$

is sometimes preferred for representing the centrum of data from skewed distributions such as the lognormal distribution. Observe that taking logarithms in Eq (12.23) establishes that the log of the geometric mean is the arithmetic mean of the log transformed data.

The *harmonic mean*,  $\bar{x}_h$ , on the other hand, defined as:

$$\frac{1}{\bar{x}_h} = \frac{1}{n} \sum_{i=1}^n \frac{1}{x_i}, \quad (12.24)$$

is more appropriate for data involving rates, ratios, or any phenomenon where the true variable of concern occurs naturally as a reciprocal entity. The classic example is data involving velocities: if a particle covers a fixed distance,  $s$ , at varying velocities  $x_1$  and  $x_2$ , from elementary physics, we are able to deduce that the average velocity with which it covers this distance is *not* the arithmetic mean  $(x_1 + x_2)/2$ , but the harmonic mean. This, of course, is because, with the distance fixed, the consequence of the variable velocity is a commensurate variation in the time to cover the distance, a reciprocal of the velocity. Note that if the *time* of travel, as opposed to the distance, is the fixed quantity, then the average velocity will be the arithmetic mean.

In general the following relationship holds between these various sample averages:

$$\bar{x}_h < \bar{x}_g < \bar{x} \quad (12.25)$$

Note how, by definition, the arithmetic mean is susceptible to undue influence from extremely large observations; by the same token, the reverse is the case with the harmonic mean which is susceptible to the undue influence of unusually small observations (whose reciprocals become unusually large). Such influences are muted with the geometric mean.

### Sample Median

Let us begin by reordering the observations  $x_1, x_2, \dots, x_n$  in ascending order to obtain  $x^{(1)}, x^{(2)}, \dots, x^{(m)}, \dots, x^{(n)}$  (we could also do this in descending order instead); if  $n$  is odd, the middle number,  $x^{(m)}$ , is the median, where  $m = (n + 1)/2$ .

If  $n$  is even, let  $m = n/2$ ; then the median is the average of the two middle numbers  $x^{(m)}$  and  $x^{(m+1)}$ , i.e.

$$x_m = \frac{x^{(m)} + x^{(m+1)}}{2} \quad (12.26)$$

Because the median, unlike the means, does not involve carrying out any arithmetic operation on the extreme values,  $x^{(1)}$  and  $x^{(n)}$ , it is much less susceptible to unusually large or unusually small observations. The median is therefore quite robust against outliers. Nevertheless, because it does not utilize all the information contained in the sample data set, it is more susceptible to chance fluctuations.

### Sample Mode

The *sample mode* can only be obtained directly from the frequency distribution.

### 12.4.3 Measures of Variability

Averages by themselves do not (and indeed cannot) adequately describe the entire data distribution: they locate the center but give no information about how the data are clustered around this center. The following are some popular measures of sample variability or spread.

#### Range

This is the simplest measure of variability or spread in a sample data set. Upon ordering the data in ascending order  $x^{(1)}, x^{(2)}, \dots, x^{(n)}$ , the sample range is simply the difference between the largest and smallest values, i.e.

$$R = x^{(n)} - x^{(1)} \quad (12.27)$$

Because it is strictly a measure of the size of the interval covered by the sample data, and does not take any other observation into consideration, it is considered a hasty measure which is very susceptible to chance fluctuations, making it somewhat unstable.

#### Average Deviation

Define the deviation of each observation  $x_i$  from the sample average,  $\bar{x}$ , as:

$$d_i = x_i - \bar{x} \quad (12.28)$$

(note that  $\sum_{i=1}^n d_i = 0$ ); then

$$\bar{d} = \frac{1}{n} \sum_{i=1}^n |d_i|, \quad (12.29)$$

known as the average deviation, provides a measure of the average absolute deviation from the mean. If the sample mean is replaced by the median  $x_m$ , then, the result is

$$\bar{d}_m = \frac{1}{n} \sum_{i=1}^n |x_i - x_m|, \quad (12.30)$$

a quantity known as the “mean absolute deviation from the median (MADM).” Because the median is more robust to outliers than the sample average, the MADM,  $\bar{d}_m$ , is, by the same token, also more robust to outliers than  $\bar{d}$ .

#### Sample Variance and Standard Deviation

The mean squared deviation from the sample mean, defined as:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2, \quad (12.31)$$

**TABLE 12.6:** Descriptive statistics for yield data sets  $Y_A$  and  $Y_B$ 

| Characteristic        | $Y_A$ values | $Y_B$ values |
|-----------------------|--------------|--------------|
| Mean                  | 75.52        | 72.47        |
| Standard Deviation    | 1.43         | 2.76         |
| Variance              | 2.05         | 7.64         |
| Skewness              | 0.32         | 0.17         |
| “Kurtosis”            | -0.09        | -0.73        |
| Total number, $n$     | 50           | 50           |
| Min                   | 72.30        | 67.33        |
| First quartile, $Q_1$ | 74.70        | 70.14        |
| Median                | 75.25        | 72.44        |
| Third quartile, $Q_3$ | 76.60        | 74.68        |
| Max                   | 79.07        | 78.49        |

is a more popular measure of variability or spread; it is the sample version of the population variance. In this context, the following is an important implication of the results of the preceding discussion on measures of central tendency (that the smallest possible mean squared deviation,  $E[(X - c)^2]$ , is achieved when  $c = \mu$ ): the smallest achievable mean squared deviation is the population variance,  $\sigma^2 = E[(X - \mu)^2]$ , the mean squared deviation from the mean; the mean squared deviation from any other value (central or not) is therefore greater than  $\sigma^2$ .

The positive square root of the sample variance,

$$s = +\sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} \quad (12.32)$$

is the sample *standard deviation*; it has the same unit as  $x$ , as opposed to  $s^2$  which has the unit of  $x$  squared.

#### Example 12.4 SUMMARY DESCRIPTIVE STATISTICS FOR YIELD DATA SETS OF CHAPTER 1

First obtain, then compare and contrast, summary descriptive statistics for the yield data sets  $Y_A$  and  $Y_B$  presented in Chapter 1 (Table 1.1).

##### Solution:

The summary descriptive statistics for these data sets, (obtainable using any typical software package), are shown in Table 12.6.

The computed *average* is higher for  $Y_A$  than for  $Y_B$ , but the *standard deviation* (hence also the *variance*) is lower for  $Y_A$ ; the very low *skewness* values for both data sets indicates a lack of asymmetry. The values shown above for “kurtosis” are actually for the so-called “excess kurtosis” defined as  $(\gamma_4 - 3)$ , which will be zero for a perfectly Gaussian distribution; the computed values shown here indicate that both data distributions are essentially Gaussian.

The remaining quantitative descriptors make up the “five-number

summary” used to produce the typical box-plot; jointly, they indicate what we already know from Fig 12.11: in every single one of these categories, the value for  $Y_A$  is consistently higher than the corresponding value for  $Y_B$ .

The modes, which cannot be computed directly from data, are obtained from the histograms (or frequency polygons) as  $y_A^* = 75$ ;  $y_B^* = 70$  for the specific bin sizes used to generate the respective histograms/frequency polygons (see Figs 12.8 and Fig 1.2 in Chapter 1; or Figs 12.9, 12.10).

This example recalls the still-unsolved problem posed in Chapter 1, and it is reasonable to ask whether the quantitative comparisons shown above are sufficient to lead us to conclude that Process A is better than Process B. While indeed these results seem to indicate that process A *might* actually be “better” than process B, i.e. that  $Y_A > Y_B$ , any such categorical statement (that  $Y_A > Y_B$ ) made at this point, strictly on the basis of this particular data set alone, will be merely speculative. The summary statistics in Table 12.6 apply only to this particular set of data; a different set of sample data from the processes will produce different data and hence different summary statistics. To make any categorical statement about which process is better and by how much requires more rigorous techniques of statistical inference that are addressed in the remaining chapters of Part IV.

#### 12.4.4 Supplementing Numerics with Graphics

It is easy to misconstrue the two aspects of descriptive statistics — graphical techniques and numerical descriptors — as mutually exclusive; or, at the very least, that the former is more useful for qualitative data while the latter is more useful for quantitative data. While there is an element of truth to this latter statement, it is more precise to consider the two aspects rather as *complementary*. Graphical techniques are great for conveying a general sense of the information contained in the data but they cannot be used for quantitative analysis or comparisons. On the other hand, even though these graphical techniques are not quantitative, they provide insight into the nature of the data set that mere numbers alone cannot possibly convey. One is incomplete without the other.

To illustrate this last point, we now present a classic example due to Anscombe<sup>3</sup>. The example involves four data sets, the first of which is shown in Table 12.7.

The basic numerical characteristics of  $X_1$  and  $Y_1$  are as follows: Total number,  $n = 11$  for both variables; the averages:  $\bar{x}_1 = 9.0$ ;  $\bar{y}_1 = 7.5$ ; the standard deviations:  $s_{x_1} = 3.32$ ;  $s_{y_1} = 2.03$ ; and the correlation coefficient

---

<sup>3</sup>Anscombe, Francis (1973), “Graphs in Statistical Analysis,” *The American Statistician*, pp. 195-199.

**TABLE 12.7:**  
The Anscombe data  
set 1

| $X_1$ | $Y_1$ |
|-------|-------|
| 10.00 | 8.04  |
| 8.00  | 6.95  |
| 13.00 | 7.58  |
| 9.00  | 8.81  |
| 11.00 | 8.33  |
| 14.00 | 9.96  |
| 6.00  | 7.24  |
| 4.00  | 4.26  |
| 12.00 | 10.84 |
| 7.00  | 4.82  |
| 5.00  | 5.68  |

**TABLE 12.8:** The Anscombe data sets  
2, 3, and 4

| $X_2$ | $Y_2$ | $X_3$ | $Y_3$ | $X_4$ | $Y_4$ |
|-------|-------|-------|-------|-------|-------|
| 10.00 | 9.14  | 10.00 | 7.46  | 8.00  | 6.58  |
| 8.00  | 8.14  | 8.00  | 6.77  | 8.00  | 5.76  |
| 13.00 | 8.74  | 13.00 | 12.74 | 8.00  | 7.71  |
| 9.00  | 8.77  | 9.00  | 7.11  | 8.00  | 8.84  |
| 11.00 | 9.26  | 11.00 | 7.81  | 8.00  | 8.47  |
| 14.00 | 8.10  | 14.00 | 8.84  | 8.00  | 7.04  |
| 6.00  | 6.13  | 6.00  | 6.08  | 8.00  | 5.25  |
| 4.00  | 3.10  | 4.00  | 5.39  | 19.00 | 12.50 |
| 12.00 | 9.13  | 12.00 | 8.15  | 8.00  | 5.56  |
| 7.00  | 7.26  | 7.00  | 6.42  | 8.00  | 7.91  |
| 5.00  | 4.74  | 5.00  | 5.73  | 8.00  | 6.89  |

between the two variables:  $\rho_{xy} = 0.82$ . A scatter plot of  $Y_1$  versus  $X_1$  is shown in Fig 12.19, which indicates a reasonably strong linearly relationship between the two variables, as correctly quantified by  $\rho_{xy} = 0.82$ .

And now, let us consider the remaining data sets 2, 3, and 4, respectively for the variables pairs  $(X_2, Y_2)$ ,  $(X_3, Y_3)$ , and  $(X_4, Y_4)$  as shown in Table 12.8. As the reader is encouraged to confirm, (see Exercise 12.17) the basic numerical characteristics of each  $(X, Y)$  pair in this data table are easily obtained as follows: Total number,  $n = 11$  for each variable; the averages:  $\bar{x}_2 = \bar{x}_3 = \bar{x}_4 = 9.0$ ;  $\bar{y}_2 = \bar{y}_3 = \bar{y}_4 = 7.5$ ; the standard deviations:  $s_{x_2} = s_{x_3} = s_{x_4} = 3.32$ ;  $s_{y_2} = s_{y_3} = s_{y_4} = 2.03$ ; and the correlation coefficient between each set of paired variables,  $\rho_{xy} = 0.82$  for  $(X_2, Y_2)$ ,  $(X_3, Y_3)$ , and  $(X_4, Y_4)$ .

Not only are these sets of characteristic numbers identical for these three data sets, they are also *identical* to the ones for the first data set,  $(X_1, Y_1)$ ;

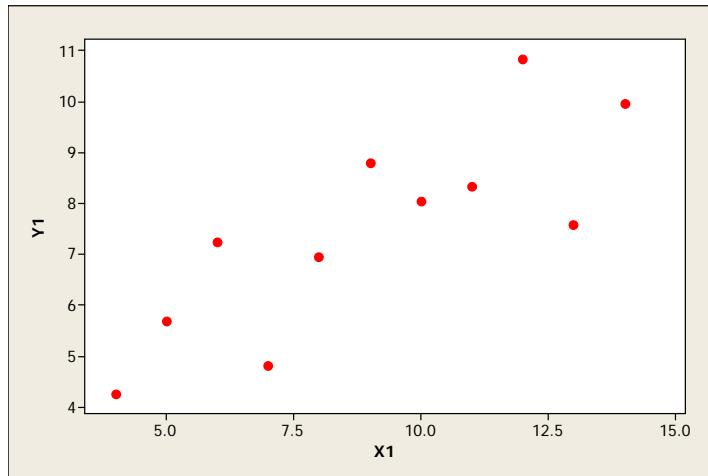
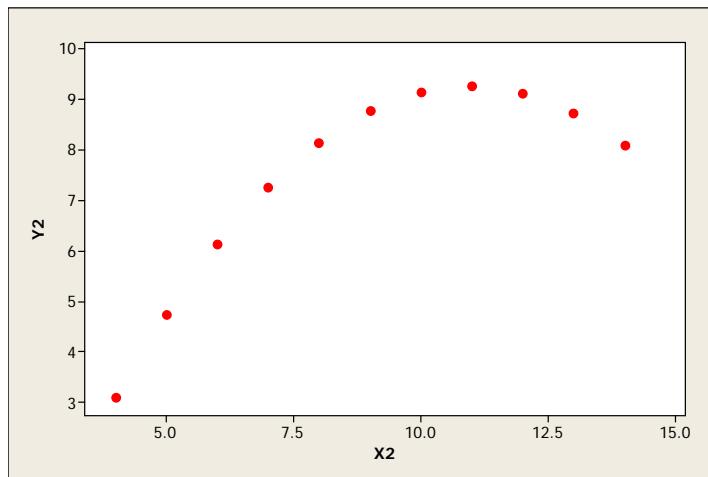
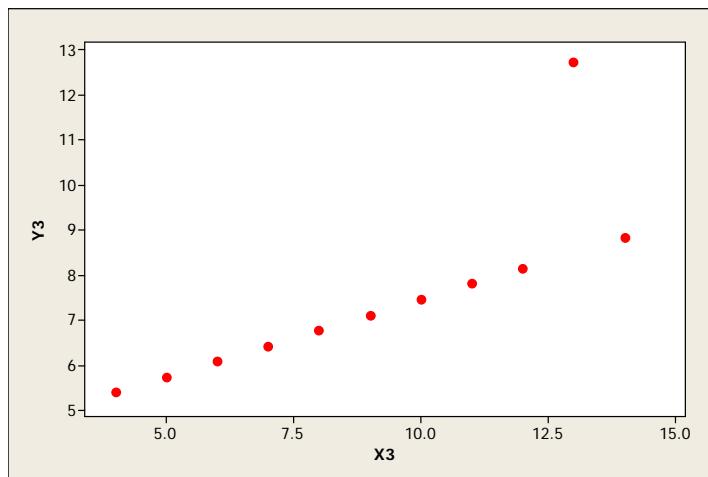


FIGURE 12.19: Scatter plot of  $Y_1$  and  $X_1$  from Anscombe data set 1.

and on this basis alone, one might be tempted to conclude that the four data sets must somehow be equivalent. Of course this is not the case. Yet, how truly different the data sets are becomes quite obvious by mere inspection of the scatter plots shown in Figs 12.20, 12.21, and 12.22 when compared to Fig 12.19.

As noted earlier, while data set 1 indicates a strong linearly relationship between the two variables (correctly implied by  $\rho_{xy} = 0.82$ ), data set 2 clearly indicates a quadratic relationship between  $Y_2$  and  $X_2$ . Data set 3, on the other hand, indicates an otherwise perfectly linear relationship that was somehow corrupted by a lone outlier, the third entry (13, 12.74). Data set 4 indicates what could best be considered as the result of a “strange” 2-level experimental design involving 10 replicates of the experiment at the low value,  $X_4 = 8$ , along with a single experiment at the high value,  $X_4 = 19$ .

Thus, while good for summarizing data with a handful of quantitative characteristics, numerical descriptions are necessarily incomplete; they can (and often) omit, or filter out, important distinguishing features in the data sets. For a complete view of the information contained in any data set, it is important to supplement quantitative descriptions with graphical representations.

FIGURE 12.20: Scatter plot of  $Y_2$  and  $X_2$  from Anscombe data set 2.FIGURE 12.21: Scatter plot of  $Y_3$  and  $X_3$  from Anscombe data set 3.

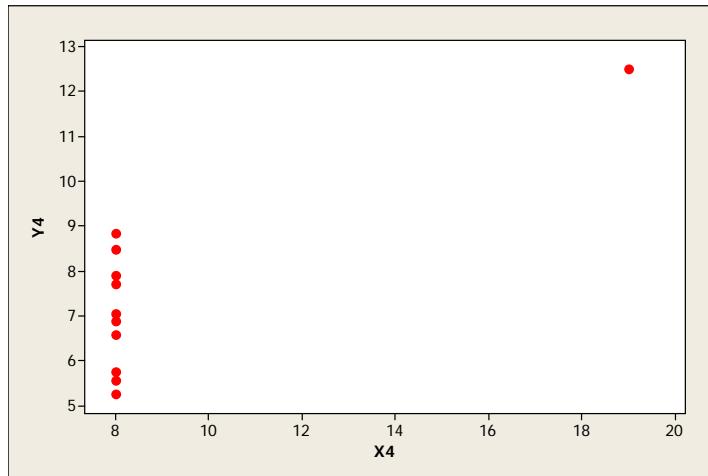


FIGURE 12.22: Scatter plot of  $Y_4$  and  $X_4$  from Anscombe data set 4.

## 12.5 Summary and Conclusions

This chapter was designed to serve as a transitional link between probability and statistics. By first articulating the central issue in statistical analysis (characterizing randomly varying phenomena on the basis of finite data) the case was made for why statistics must rely on probability even as it complements it. This led to the introduction of the basic concepts involved in statistics and an overview of what the upcoming detailed study of statistics entails. Compared to what is covered in the rest of Part IV, this chapter's introduction to descriptive statistics—organization, graphical representation and numerical summarization of sample data—may have been brief, but it is no less important. It is worth reiterating therefore that numerical analysis is most effective when complemented (wherever possible) with graphical plots.

Here are some of the main points of the chapter again.

- Statistics is concerned with fully characterizing randomly varying phenomena on the basis of finite sample data; it relies on probability to quantify the inevitable uncertainty associated with such an endeavor.
- The three central concepts of statistical analysis are:
  - *Population*: the complete collection of all the data obtainable from an experiment; unrealizable in practice, it is to statistics what the sample/random variable space is to probability;
  - *Sample*: specific observations of finite size; a subset of the population;

- *Statistical Inference*: conclusions drawn about a population from an analysis of a sample, including a measure of the associated uncertainty.
  - The three aspects of statistics are:
    - *Descriptive Statistics*: organizing, summarizing, and interpreting data;
    - *Inferential Statistics*: drawing inference from data;
    - *Statistical Design of Experiments*: systematically acquiring informative data.
- 

## APPENDIX

We wish to find the value of  $c$  that minimizes the objective function:

$$\phi_\infty = \lim_{p \rightarrow \infty} \phi_p = \lim_{p \rightarrow \infty} \int_{-\infty}^{\infty} (x - c)^p f(x) dx \quad (12.33)$$

for the continuous random variable,  $X$ . Upon taking derivatives and equating to zero, we obtain:

$$\frac{\partial \phi_\infty}{\partial c} = \lim_{p \rightarrow \infty} \left\{ -p \int_{-\infty}^{\infty} (x - c)^{p-1} f(x) dx \right\} = 0 \quad (12.34)$$

where integration by parts yields:

$$\frac{\partial \phi_\infty}{\partial c} = 0 = \lim_{p \rightarrow \infty} \left\{ -f(x)(x - c)^p \Big|_{-\infty}^{\infty} \right\} + \lim_{p \rightarrow \infty} \int_{-\infty}^{\infty} (x - c)^p f'(x) dx \quad (12.35)$$

The first term on the RHS of the equality sign after 0 vanishes because, for all valid pdfs,  $f(\infty) = f(-\infty) = 0$ , so that Eq (12.35) reduces to:

$$\frac{\partial \phi_\infty}{\partial c} = 0 = \lim_{p \rightarrow \infty} \left\{ \int_{-\infty}^{\infty} (x - c)^p f'(x) dx \right\} \quad (12.36)$$

and the indicated limit can only be zero if:

$$f'(x) = 0 \quad (12.37)$$

and this occurs at the mode of the pdf,  $f(x)$ . It is left as an exercise to the reader to obtain the corresponding result for a discrete  $X$  (See Exercise 12.19).

## REVIEW QUESTIONS

- 1.** As presented in this chapter, what are the three distinct but related entities involved in a systematic study of randomly varying phenomena?
- 2.** What is the difference between  $X$ , the random variable, and  $n$  individual observations,  $\{x_i\}_{i=1}^n$ ?
- 3.** What is the difference between writing a pdf as  $f(x)$  and as  $f(x|\theta)$ , where  $\theta$  is the vector of parameters?
- 4.** Why is the theoretical description  $f(x)$  for any specific randomly varying phenomena never completely available?
- 5.** With probability analysis, which of these two entities,  $f(x)$  and  $\{x_i\}_{i=1}^n$ , is available and what is to be determined with it? With statistical analysis, on the other hand, which of the two entities is available and what is to be determined with it?
- 6.** Statistics is a methodology for doing what?
- 7.** As stated in Section 12.1.2, what three concepts are central to statistical analysis?
- 8.** What is the difference between a sample and a population?
- 9.** What is statistical inference?
- 10.** Which of the following two entities can be specified *a-priori*, and which is an *a-posteriori* entity: (a) the random variable space,  $V_X$ , in probability, and (b) the population in statistics?
- 11.** Why must one settle for characterizing the population by drawing statistical inference?
- 12.** How does systematic data collection fit into statistical inference?
- 13.** What is the connection between probability, statistics and design of experiments?
- 14.** Into which two categories is statistics primarily categorized?
- 15.** What is “Descriptive Statistics”?
- 16.** What is “Inductive Statistics”?
- 17.** What does “Design of Experiments” entail?
- 18.** What is the difference between a qualitative and a quantitative variable?

- 19.** What is the difference between nominal and ordinal data?
- 20.** What is a bar chart and what differentiates a Pareto chart from it?
- 21.** What is a pie chart?
- 22.** What sort of information does the pie chart excel at conveying compared to the bar chart? Conversely, what sort of information does the bar chart excel at conveying compared to the pie chart?
- 23.** What is a frequency distribution and how is it related to a histogram?
- 24.** What is the relationship between the frequency distribution of a data set and the theoretical distribution of the population from which the data was obtained?
- 25.** Why should a histogram be based on an ample amount of data?
- 26.** What is the “bin size,” and why is its choice important in generating informative histograms?
- 27.** What is the “five-number” summary for a data set?
- 28.** What is a box plot?
- 29.** What are box plots particularly adept at illustrating?
- 30.** What sort of data sets are box plots better for displaying than histograms?
- 31.** What is a scatter plot and what is it most useful for?
- 32.** What are three common measures of central tendency?
- 33.** In choosing  $c$  to minimize the expected normed deviation,  $E(|X - c|^n)$ , what quantity is obtained when  $n = 1$ , or  $n = 2$ , or  $n = \infty$ ?
- 34.** What is the difference between an arithmetic mean, a geometric mean, and a harmonic mean? Under what conditions will one be preferred over the others?
- 35.** Define three common measures of variability.
- 36.** In what way do graphical techniques complement quantitative numerical techniques of summarizing data?
- 37.** What is the main lesson of the Anscombe data sets discussed in Section 12.4.4?

## EXERCISES

### Section 12.1

**12.1** Consider the experiment of tossing a single six-faced die and recording the number shown on the top face after the die comes to rest.

- (i) What is the random variable in this case, and what is the random variable space?
- (ii) What is the population and what will constitute a sample from this population?
- (iii) With adequate justification, postulate a probability model for this random variable.

**12.2** Consider a chess player who is participating in a two-game, pre-tournament qualification series where the outcome of each game is either a win, a loss, or a draw.

- (i) If we are interested only in the total number of wins *and* the total number of draws, what is the random variable in this case, its dimensionality, and the associated random variable space?
- (ii) Describe the population in this case and what will constitute a sample from such a population.
- (iii) With adequate justification, postulate a reasonable probability model for this random variable. What are the unknown parameters?

**12.3** Lucas (1985)<sup>4</sup> studied the number and frequency of occurrences of accidents over a 10-year period at a DuPont company facility. If the variable of interest is the time between occurrences of these accidents, describe the random variable space, the population, and what might be considered as a sample. Postulate a reasonable probability model for this variable and note how many parameters it has.

**12.4** In studying the useful lifetime (in years) of a brand of washing machines with the aid of a Weibull probability model, indicate the random variable space, the population and the population parameters. How can one go about obtaining a sample  $\{x_i\}_{i=1}^{50}$  from this population?

### Section 12.2

**12.5** Classify each of the following variables as either quantitative or qualitative; if quantitative, specify whether it is discrete or continuous; if qualitative whether it is ordinal or nominal.

- (i) Additive Type, ( $A$ ,  $B$ , or  $C$ ); (ii) Additive Concentration, (moles/liter); (iii) Heat Condition, (Low, Medium High); (iv) Agitation rate, (rpm); (v) Total reactor Volume, (liters); (vi) Number of reactors in ensemble, (1, 2, or 3).

**12.6** The Lucas (1985) study of Exercise 12.3 involved the following variables:

- (i) Period, (I, II); (ii) Length of Study, (Years); (iii) Number of accidents; (iv) Type of Accident; (v) Time between accidents.

Classify each variable as either quantitative or qualitative; if quantitative, specify whether it is discrete or continuous; if qualitative whether it is ordinal or nominal.

---

<sup>4</sup>Lucas J. M., (1985). "Counted Data CUSUMs," *Technometrics*, 27, 129–144

**12.7** A study of the effect of environmental cues on cell adhesion involved the following variables. Classify each one as either quantitative (discrete or continuous) or qualitative (ordinal or nominal).

- (i) Type of stimulus, (Mechanical, Chemical, Topological); (ii) Ligand concentration;
- (iii) Surface type (Patterned, Plain); (iv) Mechanical force; (v) Number of integrin clusters; (vi) Integrin cluster size; (vii) Cell Status (Adherent, Non-Adherent).

### Section 12.3

**12.8** The table below shows where chemical engineers found employment in 2000, categorized by degree. For each degree category, (BS, MS and PhD) draw a bar chart and a pie chart. Comment on what stands out most prominently in each case within each degree category and across the degree categories (for example, “Academia” across the categories.)

| Employer                 | BS          | MS          | PhD         |
|--------------------------|-------------|-------------|-------------|
|                          | Placement % | Placement % | Placement % |
| Industry                 | 55.9        | 44.1        | 57.8        |
| Government               | 1.7         | 2.2         | 0.8         |
| Grad/Professional School | 11.2        | 33.1        | 13.1        |
| Returned to Home Country | 1.3         | 4.7         | 0.1         |
| Unemployed               | 9.5         | 4.5         | 2.8         |
| Unknown Employment       | 18.8        | 7.4         | 6.4         |
| Academia                 | 0.0         | 1.8         | 16.5        |
| Other                    | 1.8         | 2.2         | 1.7         |

**12.9** Generate Pareto Charts for each degree category for the chemical engineering employment data in Exercise 12.8. Interpret these charts.

**12.10** The data in the table below (adapted from a 1983 report from the President’s council on Physical Fitness and Sports) shows the number of adult Americans (non-professionals) participating in the indicated sports.

| Type<br>of<br>Sport | Number of<br>Participants<br>(in millions) |
|---------------------|--|
| Basketball          | 29   |
| Bicycling           | 44   |
| Football (Touch)    | 18   |
| Golf                | 13   |
| Hiking              | 34   |
| Ice-skating         | 11   |
| Racquetball         | 10   |
| Roller-skating      | 20   |
| Running             | 20   |
| Skiing              | 10   |
| Softball            | 26   |
| Swimming            | 60   |
| Tennis              | 23   |
| Volleyball          | 21   |

Generate a bar chart and a Pareto chart for this data; interpret the charts. Why

is it unadvisable to use a pie chart to represent such data?

**12.11** The following data set has been adapted from information provided by the IRS in 1985 about the population of “well-to-do” (WTD) individuals in various states. (At the time, a WTD individual was defined as someone with gross assets  $\geq \$500,000$ ).

| State        | Population | WTD        |       |
|--------------|------------|------------|-------|
|              |            | California | Texas |
| California   | 301,500    |            |       |
| Texas        | 204,800    |            |       |
| Florida      | 151,800    |            |       |
| New York     | 110,100    |            |       |
| Illinois     | 108,000    |            |       |
| Pennsylvania | 86,800     |            |       |
| Ohio         | 52,500     |            |       |
| New Jersey   | 51,300     |            |       |
| Iowa         | 50,800     |            |       |
| Michigan     | 48,100     |            |       |

Generate a bar chart for the data in terms of relative frequency. In how many of these 10 states will one find approximately 80% of the listed “well-to-do” individuals?

### Section 12.4

**12.12** The data in the following table shows samples of size  $n = 20$  drawn from four different populations coded as  $N, L, G$  and  $I$ . Generate a histogram and a box plot for each of the data sets. Discuss what these plots indicate about the general characteristics of the population from which the data were obtained.

| $X_N$   | $X_L$   | $X_G$   | $X_I$    |
|---------|---------|---------|----------|
| 9.3745  | 7.9128  | 10.0896 | 0.084029 |
| 8.8632  | 5.9166  | 15.7336 | 0.174586 |
| 11.4943 | 4.5327  | 15.0422 | 0.130492 |
| 9.5733  | 33.2631 | 5.5482  | 0.115567 |
| 9.1542  | 24.1327 | 18.0393 | 0.187260 |
| 9.0992  | 5.4151  | 17.9543 | 0.100054 |
| 10.2631 | 16.9556 | 12.5549 | 0.101405 |
| 9.8737  | 3.9345  | 9.6640  | 0.100835 |
| 7.8192  | 35.0376 | 14.2975 | 0.097173 |
| 10.4691 | 25.1182 | 4.2599  | 0.141233 |
| 9.6981  | 1.1804  | 19.1084 | 0.060470 |
| 10.5911 | 2.3503  | 7.0735  | 0.127663 |
| 11.6526 | 15.6894 | 7.6392  | 0.074183 |
| 10.4502 | 5.8929  | 14.1899 | 0.086606 |
| 10.0772 | 8.0254  | 13.8996 | 0.084915 |
| 10.2932 | 16.1482 | 9.7680  | 0.242657 |
| 11.7755 | 0.6848  | 8.5779  | 0.052291 |
| 9.3790  | 6.6974  | 7.5486  | 0.116172 |
| 9.9202  | 3.6909  | 10.4043 | 0.084339 |
| 10.9067 | 34.2152 | 14.8254 | 0.205748 |

**12.13** For each sample in Exercise 12.12, compute the (i) arithmetic mean; (ii) geometric mean; (iii) median; and (iv) harmonic mean. Which do you think is a more

appropriate measure of the central tendency of the original population from which these samples were drawn, and why?

**12.14** The table below shows a relative frequency summary of sample data on distances between DNA replication origins (inter-origin distances), measured by Li et al., (2003)<sup>5</sup>, with an *in vitro* Xenopus egg extract assay in Chinese Hamster Ovary (CHO) cells, as reported in Chapter 7 of Birtwistle (2008)<sup>6</sup>.

| Inter-Origin Distance (kb) | Relative Frequency<br>$f_r(x)$ |
|----------------------------|--------------------------------|
| 0                          | 0.00                           |
| 15                         | 0.09                           |
| 30                         | 0.18                           |
| 45                         | 0.26                           |
| 60                         | 0.18                           |
| 75                         | 0.09                           |
| 90                         | 0.04                           |
| 105                        | 0.03                           |
| 120                        | 0.05                           |
| 135                        | 0.04                           |
| 150                        | 0.01                           |
| 165                        | 0.02                           |

(The data set is similar to, but different from, the one in Application Problem 9.40 in Chapter 9.) Obtain a histogram of the data and determine the mean, variance and your best estimate of the median.

**12.15** From the data given in Table 12.5 in the text, generate a scatter plot of (i) city gas mileage against engine capacity, and (ii) city gas mileage against number of cylinders, for the two-seater automobiles listed in that table. Compare these plots to the corresponding ones in the text for highway gas mileage. Are there any surprises in these city gas mileage plots?

**12.16** Let  $X_1$  and  $X_2$  represent, respectively, the engine capacity, in liters, and number of cylinders for the population of two-seater automobiles; let  $Y_1$  and  $Y_2$  represent the city gas mileage and highway gas mileage, respectively, for these same automobiles. Consider that the data in Table 12.5 constitute appropriate samples from the respective populations. From the supplied sample data, compute the complete set of 6 pairwise correlation coefficients between these variables. Comment on what these correlation coefficients mean.

**12.17** Confirm that the basic numerical characteristics of each  $(X, Y)$  pair in Table 12.8 are as given in the text.

**12.18** Determine the value of  $c$  that minimizes the mean absolute deviation  $\phi_1$

<sup>5</sup>Li, F., Chen, J., Solessio, E. and Gilbert, D. M. (2003). "Spatial distribution and specification of mammalian replication origins during G1 phase." *J Cell Biol* 161, 257-66.

<sup>6</sup>M. R. Birtwistle, (2008). *Modeling and Analysis of the ErbB Signaling Network: From Single Cells to Tumorigenesis*, PhD Dissertation, University of Delaware.

between it and the possible values of the discrete random variable,  $X$ , whose pdf is given as  $f(x_i)$ , i.e.,

$$\phi_1 = \sum_{i=0}^{\infty} |x_i - c| f(x_i) dx \quad (12.38)$$

**12.19** Find the value of  $c$  that minimizes the objective function:

$$\phi_{\infty} = \lim_{p \rightarrow \infty} \phi_p = \lim_{p \rightarrow \infty} \sum_{i=0}^{\infty} (x_i - c)^p f(x_i) \quad (12.39)$$

and show that it is the mode of the discrete pdf  $f(x_i)$ , i.e.,  $f(c) > f(x_i)$  for all  $i$ .

## APPLICATION PROBLEMS

**12.20** A quality control engineer at a semi-conductor manufacturing site is concerned about the number of contaminant particles (flaws) found on each standard size silicon wafer produced at the site. A sample of 20 silicon wafers selected and examined for flaws produced the result (the number of flaws found on each wafer) shown in the following table.

|   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|
| 3 | 0 | 0 | 2 | 3 | 0 | 3 | 2 | 1 | 2 |
| 4 | 1 | 2 | 3 | 2 | 1 | 2 | 4 | 0 | 1 |

- (i) For this particular problem, what is the random variable,  $X$ , the set  $\{x_i\}_{i=1}^n$ , and why is the Poisson model, with the single parameter  $\lambda$ , a reasonable probability model for the implied phenomenon?
- (ii) From the expression for  $f(x|\lambda)$ , compute the theoretical probabilities when the population parameter is specified as 0.5, 1.0 and 1.5. From these theoretical probabilities, which of the postulated population parameters appears more representative of observations?

**12.21** The time in months between occurrences of safety violations in a toll manufacturing facility is shown in the table below.

|      |      |      |      |      |      |      |      |      |      |
|------|------|------|------|------|------|------|------|------|------|
| 1.31 | 0.15 | 3.02 | 3.17 | 4.84 | 0.71 | 0.70 | 1.41 | 2.68 | 0.68 |
| 1.94 | 3.21 | 2.91 | 1.66 | 1.51 | 0.30 | 0.05 | 1.62 | 6.75 | 1.29 |
| 0.79 | 1.22 | 0.65 | 3.90 | 0.18 | 0.57 | 7.26 | 0.43 | 0.96 | 3.76 |

- (i) Determine the mean, median and variance for this sample data. Construct a histogram and explain why the observed shape is not surprising, given the nature of the phenomenon in question.
- (ii) What is a reasonable probability model for the population from which the data came? If the population parameter, the mean time between violations, is postulated to be 2 months, compute the theoretical probability of going more than 2 months without a safety violation. Is this theoretical probability compatible with this sample data? Explain.
- (iii) In actual fact, the data were obtained for three different operators and have been

arranged accordingly: the first row is for “Operator *A*,” the second row, for “Operator *B*,” and the third row for “Operator *C*.” It has been a long-held preconception in the manufacturing facility that “Operator *A*” is relatively more safety-conscious than the other two. Strictly on the basis of any graphical and numerical descriptions of each data set that you deem appropriate, is there any “suggestion” in this data set that could potentially support this preconception? Explain.

**12.22** Nelson (1989)<sup>7</sup> quantified the “cold cranking power” of five different battery types in terms of the number of seconds that a particular battery generated its rated amperage without falling below 7.2 volts, at 0° F. The experiment was repeated four times for each battery type and the resulting data set is shown in the following table.

|                |   | 1               | 2  | 3  | 4  | 5  |
|----------------|---|-----------------|----|----|----|----|
|                |   | Experiment No ↓ |    |    |    |    |
| Battery Type → |   | 1               | 2  | 3  | 4  | 5  |
|                | 1 | 41              | 42 | 27 | 48 | 28 |
|                | 2 | 43              | 43 | 26 | 45 | 32 |
|                | 3 | 42              | 46 | 28 | 51 | 37 |
|                | 4 | 46              | 38 | 27 | 46 | 25 |

Is there any suggestion of “descriptive” (as opposed to “inductive”) evidence in this data set to support the postulate that some battery types are better than others? Which ones appear better? (More precise “inductive” methods for answering such questions are discussed in Chapters 15 and 19.)

**12.23** Consider the data in the following table showing the boiling point (in °C) of 8 hydrocarbons in a homologous series, along with *n*, the number of carbon atoms in each molecule.

| Hydrocarbon Compound | <i>n</i> , Number of Carbon Atoms | Boiling Point °C |
|----------------------|-----------------------------------|------------------|
| Methane              | 1                                 | -162             |
| Ethane               | 2                                 | -88              |
| Propane              | 3                                 | -42              |
| n-Butane             | 4                                 | 1                |
| n-Pentane            | 5                                 | 36               |
| n-Hexane             | 6                                 | 69               |
| n-Heptane            | 7                                 | 98               |
| n-Octane             | 8                                 | 126              |

What does this data set imply about the possibility of predicting the boiling point of compounds in this series on the basis of the number of carbon atoms? Compute the correlation coefficient between these two variables, even though the number *n* is not a random variable. Comment on what the computed value indicates.

**12.24** The following table contains experimental data on the thermal conductivity, *k* (W/m·°C), of a metal, determined at various temperatures.

<sup>7</sup>Nelson, L.S., (1989). “Comparison of Poisson means,” *J. of Qual. Tech.*, 19, 173–179.

| $k$ (W/m·°C) | Temperature (°C) |
|--------------|------------------|
| 93.228       | 100              |
| 92.563       | 150              |
| 99.409       | 200              |
| 101.590      | 250              |
| 111.535      | 300              |
| 115.874      | 350              |
| 119.390      | 400              |
| 126.615      | 450              |

What sort of systematic functional relationship between the two variables (if any) does the evidence in the data suggest? Compute the correlation coefficient between the two variables and comment on what this value indicates. What would you recommend as a reasonable value to postulate for the thermal conductivity of the metal at 325 °C? Justify your answer succinctly.

**12.25** The following data set, from the same study by Lucas (1985) referenced in Exercise 12.3, shows the actual number of accidents occurring per quarter (three months) separated into two periods: Period I is the first five-year period of the study; Period II, the second five-year period.

| Period I |   |    |    | Period II |   |   |   |
|----------|---|----|----|-----------|---|---|---|
| 5        | 5 | 10 | 8  | 3         | 4 | 2 | 0 |
| 4        | 5 | 7  | 3  | 1         | 3 | 2 | 2 |
| 2        | 8 | 6  | 9  | 7         | 7 | 1 | 4 |
| 5        | 6 | 5  | 10 | 1         | 2 | 2 | 1 |
| 6        | 3 | 3  | 10 | 4         | 4 | 4 | 4 |

Provide an appropriate statistical description and summary of this data set; comment on any distinctive characteristics and postulate potential explanations for your observation.

**12.26** According to census records, the age distribution of the inhabitants of the United States in 1960 and in 1980 is as shown in the table below.

| Age Group | 1960   | 1980   |
|-----------|--------|--------|
| < 5       | 20,321 | 16,348 |
| 5–9       | 18,692 | 16,700 |
| 10–14     | 16,773 | 18,242 |
| 15–19     | 13,219 | 21,168 |
| 20–24     | 10,801 | 21,319 |
| 25–29     | 10,869 | 19,521 |
| 30–34     | 11,949 | 17,561 |
| 35–39     | 12,481 | 13,965 |
| 40–44     | 11,600 | 11,669 |
| 45–49     | 10,879 | 11,090 |
| 50–54     | 9,606  | 11,710 |
| 55–59     | 8,430  | 11,615 |
| 60–64     | 7,142  | 10,088 |
| ≥ 65      | 16,560 | 25,550 |

From an appropriate statistical description of these data sets, comment on the indicated changes in the population in the two decades between 1960 and 1980.

Identify any features that might be due to the “baby-boom” generation—those born during the period from the end of World War II until about 1965.



# Chapter 13

---

## Sampling

|        |   |     |
|--------|---|-----|
| 13.1   | Introductory Concepts .....                             | 459 |
| 13.1.1 | The Random Sample .....                                 | 460 |
| 13.1.2 | The “Statistic” and its Distribution .....              | 461 |
|        | Definitions .....                                       | 461 |
|        | Utility .....   | 462 |
| 13.2   | The Distribution of Functions of Random Variables ..... | 463 |
| 13.2.1 | General Overview .....                                  | 463 |
| 13.2.2 | Some Important Sampling Distribution Results .....      | 463 |
| 13.3   | Sampling Distribution of The Mean .....                 | 465 |
| 13.3.1 | Underlying Probability Distribution Known .....         | 465 |
| 13.3.2 | Underlying Probability Distribution Unknown .....       | 467 |
| 13.3.3 | Limiting Distribution of the Mean .....                 | 467 |
| 13.3.4 | $\sigma$ Unknown .....                                  | 470 |
| 13.4   | Sampling Distribution of the Variance .....             | 472 |
| 13.5   | Summary and Conclusions .....                           | 476 |
|        | REVIEW QUESTIONS .....                                  | 477 |
|        | EXERCISES .....   | 478 |
|        | APPLICATION PROBLEMS .....                              | 482 |

*If in other sciences we should arrive  
at certainty without doubt and truth without error,  
it behooves us to place the foundation of knowledge in mathematics.*

Roger Bacon (c.1220–c.1292) *Opus Majus, Bk.1 Ch. 4*

If, as stated in Chapter 12, inductive (or inferential) statistics is primarily concerned with drawing inference about a population from sample information, then a logical treatment of inductive statistics must begin with *sampling* — a formal study of samples from a population. Because it is a finite collection of individual observations, a sample is itself susceptible to random variation since different samples drawn from the same population under identical conditions will be different. As such, before samples can be useful for statistical inference concerning the populations that produced them, the variability inherent in samples must be characterized mathematically (just as was done for individual observations  $x_i$  from a random variable,  $X$ ). How one characterizes the variability inherent in samples, as distinct from, but obviously related to, characterizing the variability inherent in individual observations of a random variable,  $X$ , is the focus in this chapter. Sampling is the foundational element of statistical inference, and this chapter’s discussion is an indispensable precursor to the discussions of estimation and hypothesis testing to follow in the next two chapters.

### 13.1 Introductory Concepts

As we now know, the role played by the sample space (or, equivalently, the random variable space) in probability theory is analogous to that of the population in statistics. In this regard, what the randomly varying individual observation,  $x$ , is to the random variable space,  $V_X$ , in probability theory, the finite-sized sample,  $\{x_i\}_{i=1}^n$ , is to the population in statistics. In the former, the variability inherent to individual observations is characterized by the pdf,  $f(x)$ , an ensemble representation that is then used to carry out theoretical probabilistic analysis for the elements of  $V_X$ . There is an analogous problem in statistics: in order to characterize the population appropriately, we must first figure out how to characterize the variability intrinsic to the finite-sized sample. The entire subject matter of characterizing and analyzing samples from a population, and employing such results to make statistical inference statements about the population, is known as sampling, or sampling theory.

The following three concepts,

1. The Random Sample;
2. The “Statistic”; and
3. The Distribution of a “Statistic” (or The “Sampling Distribution”),

are central to sampling. As discussed in detail shortly, sampling theory combines these concepts into the basis for characterizing the uncertainty in samples in terms of probability distributions that can then be used for statistical inference.

#### 13.1.1 The Random Sample

Since the finite-sized sample is the only source of information about an entire population, it is essential that the sample be representative of the population. This is the primary motivation behind the concept of the *random sample*, explained as follows. Consider a set of observations (data)  $\{x_1, x_2, \dots, x_n\}$  drawn from a population of size  $N$  (where  $N$  is possibly infinite): if all possible subsets  $n$  of the  $N$  elements of the population have equal probabilities of being chosen, then the observations constitute a random sample from the population. The rationale is clear: with equal probabilities of selection, no particular subset will preferentially favor any particular aspect of the population. The mathematical implication of this concept (sometimes considered as the formal definition) now follows.

**Definition:** Let  $X_1, X_2, \dots, X_n$  denote  $n$  mutually, stochastically, independent random variables, each of which has the same, but possibly unknown, pdf  $f(x)$ ; i.e. the pdfs of  $X_1, X_2, \dots, X_n$  are, respectively,  $f_1(x_1) = f(x_1); f_2(x_2) = f(x_2); \dots; f_n(x_n) = f(x_n)$ , so that the joint pdf is  $f(x_1)f(x_2)\cdots f(x_n)$ . The random variables  $X_1, X_2, \dots, X_n$  constitute a random sample from a distribution that has the pdf  $f(x)$ .

The condition noted here for the random variables is also sometimes rendered as “independently and identically distributed” or i.i.d.

The practical implication of this concept and the definition given above is that if we can ensure that the sample from a population is drawn “randomly,” then the joint distribution of the sample is a product of the contributing pdfs. This rather simple concept significantly simplifies the theory and practice of statistical inference, as we show shortly.

### 13.1.2 The “Statistic” and its Distribution

#### Definitions

A “statistic” is any function of one or more random variables that does not depend upon any unknown population parameters. For example, let  $X_1, X_2, \dots, X_n$  be mutually stochastically independent random variables, each with identical  $N(\mu, \sigma^2)$  distributions, with unknown parameters  $\mu$  and  $\sigma$ ; then the random variable  $Y$  defined as:

$$Y = \sum_{i=1}^n X_i \quad (13.1)$$

is a *statistic*. It is a function of  $n$  random variables and it does not depend on any of the unknown parameters,  $\mu$  and  $\sigma$ . On the other hand,

$$Z = \frac{X_1 - \mu}{\sigma} \quad (13.2)$$

is a function of the random variable  $X_1$ , but it depends on  $\mu$  and  $\sigma$ ; unless these parameters are known,  $Z$  does not qualify as a statistic.

In general, for a set of random variables  $X_1, X_2, \dots, X_n$  constituting a random sample, the random variable  $Y$  that is a function of these random variables, i.e.

$$Y = g(X_1, X_2, \dots, X_n) \quad (13.3)$$

is a statistic so long as  $g(\cdot)$  is independent of unknown parameters. Observe that given the joint pdf  $f(x_1, x_2, \dots, x_n)$ , we can use Eq (13.3) to obtain  $f_Y(y)$ , the pdf of the statistic,  $Y$ . It is important to note:

1. Even though a statistic (say  $Y$  as defined above) does not depend upon an unknown parameter, the *distribution* of a statistic (say  $f_Y(y)$ ) quite often depends on unknown parameters.
2. The distributions of such statistics are called *sampling distributions* because they are distributions of functions of samples. Since a statistic, as defined above, is itself a random variable, its sampling distribution describes the variability (chance fluctuations) one will observe in it as a result of random sampling.

### Utility

The primary utility of the statistic and its distribution is in determining unknown population parameters from samples, and in quantifying the inherent variability. This becomes evident from the following re-statement of the problem of statistical inference:

1. The pdf for characterizing the random variable,  $X$ ,  $f(x|\theta)$ , contains unknown parameters,  $\theta$ ; were it possible to observe, via experimentation, the *complete* population in its entirety, we would be able to construct, from such observations, the complete  $f(x|\theta)$ , including the parameters; however, only a finite sample from the population is available via experimentation. When the form of  $f(x|\theta)$  is known we are left with the issue of determining the unknown parameters,  $\theta$ , from the sample, i.e., making inference about the population parameter  $\theta$  from sample data.
2. We make these inferences by investigating random samples, using appropriate “statistics” (quantities calculated from the random sample) that will provide information about the parameters.
3. These “statistics,” which enable us to determine the unknown parameters, are themselves random variables; the distribution of such statistics then enable us to make probability statements about these statistics and hence the unknown parameters.

It turns out that most of the unknown parameters in a pdf representing a population are “contained” in the mean,  $\mu$ , and variance,  $\sigma^2$ , of the pdf in question. Thus, once the mean and variance of a pdf are known, the naturally occurring parameters can then be deduced. For example, if  $X \sim N(\mu, \sigma^2)$ , the mean and the variance are in fact the naturally occurring parameters; for the gamma random variable,  $X \sim \gamma(\alpha, \beta)$ , recall that

$$\mu = \alpha\beta \quad (13.4)$$

$$\sigma^2 = \alpha\beta^2 \quad (13.5)$$

a pair of equations that can be solved simultaneously to yield:

$$\alpha = \mu^2/\sigma^2 \quad (13.6)$$

$$\beta = \sigma^2/\mu \quad (13.7)$$

For the Poisson random variable,  $X \sim \mathcal{P}(\lambda)$ ,  $\mu = \sigma^2 = \lambda$ , so that the parameter  $\lambda$  is directly determinable from either the mean or the variance (or both).

Thus, it is often sufficient to use statistics that represent the mean and the variance of a population to determine the unknown population parameters. It is therefore customary for sampling theory to concentrate on the sampling distributions of the mean and of the variance. And now, because statistics are functions of random variables, determining sampling distributions requires techniques for obtaining distributions of functions of random variables.

## 13.2 The Distribution of Functions of Random Variables

The general problem of interest may be stated as follows: given the joint pdf for  $n$  random variables  $X_1, X_2, \dots, X_n$ , find the pdf  $f_Y(y)$  for the random variable  $Y$  defined as

$$Y = g(X_1, X_2, \dots, X_n). \quad (13.8)$$

### 13.2.1 General Overview

This is precisely the problem discussed in some detail in Chapter 6, where various methods of solution were presented. For example, it is shown in Example 6.7 that if  $X_1 \sim \gamma(\alpha, 1)$  and  $X_2 \sim \gamma(\beta, 1)$ , then the functions defined as:

$$Y_1 = X_1 + X_2 \quad (13.9)$$

$$Y_2 = \frac{X_1}{X_1 + X_2} \quad (13.10)$$

have the following distributions:  $Y_1 \sim \gamma(\alpha + \beta, 1)$  and  $Y_2 \sim B(\alpha, \beta)$ . Also, given two Poisson random variables  $X_1 \sim \mathcal{P}(\lambda_1)$  and  $X_2 \sim \mathcal{P}(\lambda_2)$ , then a statistic defined as:

$$Y = X_1 + X_2 \quad (13.11)$$

can be shown (using methods of characteristic functions, for example) to possess a Poisson distribution with parameters  $(\lambda_1 + \lambda_2)$ , i.e.  $Y \sim \mathcal{P}(\lambda_1 + \lambda_2)$ . These general ideas can be applied specifically to sampling distributions that are of interest in statistical inference.

### 13.2.2 Some Important Sampling Distribution Results

As we will soon see, many (but not all) classical statistical inference problems involve sampling from distributions that are either exactly Gaussian or

approximately so. The following is a collection of some key results regarding sampling from the Gaussian and related distributions.

**1. Linear Combination of Gaussian Random Variables:** Consider  $n$  mutually stochastically independent random variables,  $X_1, X_2, \dots, X_n$ , with respective pdfs  $N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2), \dots, N(\mu_n, \sigma_n^2)$ ; the random variable:

$$Y = k_1 X_1 + k_2 X_2 + \dots + k_n X_n \quad (13.12)$$

where  $k_1, k_2, \dots, k_n$  are real constants, possesses a Gaussian distribution  $N(\mu_y, \sigma_y^2)$  where

$$\mu_y = k_1 \mu_1 + k_2 \mu_2 + \dots + k_n \mu_n \quad (13.13)$$

$$\sigma_y^2 = k_1^2 \sigma_1^2 + k_2^2 \sigma_2^2 + \dots + k_n^2 \sigma_n^2 \quad (13.14)$$

These results are straightforward to establish (see Exercise 13.4). In particular, if  $\mu_1 = \mu_2 = \dots = \mu_n = \mu$ , and  $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_n^2 = \sigma^2$ , so that the random variables  $X_i; i = 1, 2, \dots, n$  are all identically distributed, then:

$$\mu_y = \left( \sum_{i=1}^n k_i \right) \mu \quad (13.15)$$

$$\sigma_y^2 = \left( \sum_{i=1}^n k_i^2 \right) \sigma^2 \quad (13.16)$$

Furthermore, if  $k_i = 1/n$  for all  $i$ , then

$$\mu_y = \mu \quad (13.17)$$

$$\sigma_y^2 = \sigma^2/n \quad (13.18)$$

Even if the distributions of  $X_i$  are *not* Gaussian, but the means and variances are still  $\mu_i$  and  $\sigma_i^2$  respectively, clearly, the resulting distribution of  $Y$  will be determined by the underlying distributions of  $X_i$ , but its mean and variance,  $\mu_y$  and  $\sigma_y^2$ , will *still* be as given in Eqs (13.17) and (13.18). These results are also fairly straightforward to establish (see See Excercise 13.5).

**2. Sum of Squares of Standard Normal Variables:** Consider a random sample of size  $n$  from a Gaussian  $N(\mu, \sigma^2)$  distribution,  $X_1, X_2, \dots, X_n$ ; the random variable

$$Y = \sum_{i=1}^n \left( \frac{X_i - \mu}{\sigma} \right)^2 \quad (13.19)$$

has a  $\chi^2(n)$  distribution.

**3. Sum of Chi-Square Random Variables:** Consider  $n$  mutually stochastically independent random variables,  $X_1, X_2, \dots, X_n$ , with respective pdfs  $\chi^2(r_1), \chi^2(r_2), \dots, \chi^2(r_n)$ ; the random variable

$$Y = X_1 + X_2 + \dots + X_n \quad (13.20)$$

has a  $\chi^2(r)$  distribution with degrees of freedom,

$$r = r_1 + r_2 + \cdots + r_n. \quad (13.21)$$

These results find significant application in classic statistical inference.

---

### 13.3 Sampling Distribution of The Mean

The general problem of interest is as follows: given  $(X_1, X_2, \dots, X_n)$ , a random sample from a distribution with pdf  $f(x)$ , the statistic defined as

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (13.22)$$

is a random variable whose specific value, the actual sample average,  $\bar{x}$ , will vary from one specific random sample to another. *What is the theoretical sampling distribution of the random variable,  $\bar{X}$ ?*

As might be expected, the answer to this question depends on what is known about the distribution from which the random sample is drawn. We now discuss the cases most relevant to statistical inference.

#### 13.3.1 Underlying Probability Distribution Known

If we know  $f(x)$ , the distribution of the population from which the sample is drawn, we can use the techniques discussed in Chapter 6 (and mentioned above) to obtain the sampling distribution of the mean  $\bar{X}$ . The next two examples illustrate this point for the Gaussian pdf and the exponential pdf.

**Example 13.1: SAMPLING DISTRIBUTION: MEAN OF RANDOM SAMPLE FROM GAUSSIAN DISTRIBUTION**

If  $X_1, X_2, \dots, X_n$  is a random sample from a Gaussian distribution with mean  $\mu$  and variance  $\sigma^2$ , find the distribution of  $\bar{X}$  defined in Eq 14.69.

**Solution:**

First,  $X_1, X_2, \dots, X_n$  is a random sample from the same Gaussian distribution, whose characteristic function is:

$$\varphi(t) = \exp \left\{ j\mu t - \frac{1}{2}\sigma^2 t^2 \right\} \quad (13.23)$$

By virtue of the independence of the random variables, and employing results about the characteristic function of linear sums of independent random variables, then the characteristic function of  $\bar{X}$  is obtained as

$$\varphi_{\bar{X}}(t) = \prod_{i=1}^n \exp \left\{ \frac{j\mu}{n} t - \frac{1}{2} \frac{\sigma^2}{n^2} t^2 \right\} \quad (13.24)$$

This product of  $n$  identical exponentials becomes an exponential of the sums of the terms in the winged brackets, and finally simplifies to give:

$$\varphi_{\bar{X}}(t) = \exp \left\{ j\mu t - \frac{1}{2} \frac{\sigma^2}{n} t^2 \right\} \quad (13.25)$$

which we recognize immediately as the characteristic function of a Gaussian random variable with mean  $\mu$ , and variance  $\sigma^2/n$ . We therefore conclude that, in this case,  $\bar{X} \sim N(\mu, \sigma^2/n)$ , i.e. that the sampling distribution of the mean of a random sample of size  $n$  from a  $N(\mu, \sigma^2)$  is also a Gaussian distribution with the same mean, but with variance  $\sigma^2/n$ .

**Example 13.2: SAMPLING DISTRIBUTION: MEAN OF RANDOM SAMPLE FROM EXPONENTIAL DISTRIBUTION**

If  $X_1, X_2, \dots, X_n$  is a random sample from an exponential distribution,  $\mathcal{E}(\beta)$ , find the distribution of  $\bar{X}$  defined in Eq 14.69.

**Solution:**

Again, as with Example 13.1, since  $(X_1, X_2, \dots, X_n)$  is a random sample from the same exponential distribution, we begin by recalling the characteristic function for the exponential random variable:

$$\varphi(t) = \frac{1}{(1 - j\beta t)} \quad (13.26)$$

By virtue of the independence of the random variables, and employing results about the characteristic function of linear sums of independent random variables, the characteristic function of  $\bar{X}$  is obtained as

$$\varphi_{\bar{X}}(t) = \prod_{i=1}^n \frac{1}{(1 - j\beta \frac{t}{n})} = \frac{1}{(1 - j\frac{\beta}{n}t)^n} \quad (13.27)$$

This is recognizable as the characteristic function of a gamma random variable with parameters  $\alpha^*$  and  $\beta^*$ , where

$$\alpha^* = n \quad (13.28)$$

$$\beta^* = \beta/n \quad (13.29)$$

i.e.  $\bar{X} \sim \gamma(n, \beta/n)$ . Observe that because the mean value for a  $\gamma(\alpha^*, \beta^*)$  random variable is  $\alpha^*\beta^*$ , and the variance is  $\alpha^*\beta^{*2}$ , the implication here is that the expected value of  $\bar{X}$  in this case is  $n\beta/n = \beta$ ; and the variance is  $\beta^2/n$ .

Some important points to note from these examples:

1. If  $\mu_X$  is the expected value, and  $\sigma_X^2$  the variance, of the pdf from which the random sample was drawn, these examples show that, for both the Gaussian pdf and the exponential pdf, the expected value and variance of the sample mean,  $\bar{X}$ , are given by:

$$\mu_{\bar{X}} = \mu_X \quad (13.30)$$

$$\sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n} \quad (13.31)$$

2. What is true for these two example random variables is true in general, regardless of the underlying distribution (although we have not proved this formally).
3. The implications are as follows: the expectation of the sample mean is identical to population mean; and the variance of the sample mean goes to zero as  $n \rightarrow \infty$ . In anticipation of a more detailed discussion in Chapter 14, we note that the sample mean appears to have desirable properties that recommend its use in determining the true value of the unknown population mean.

### 13.3.2 Underlying Probability Distribution Unknown

If the form of the pdf,  $f(x)$ , underlying a population is unknown, we cannot, in general, determine the full sampling distribution for any random sample drawn from such a population. Nevertheless, the following information is still available, regardless of the underlying pdf:

If the random sample  $(X_1, X_2, \dots, X_n)$  comes from a population with mean  $\mu$  and variance  $\sigma^2$ , but whose full pdf is unknown, then the sample mean  $\bar{X}$  is a random variable whose mean  $\mu_{\bar{X}}$  and variance  $\sigma_{\bar{X}}^2$  are given by the following expressions:

$$\mu_{\bar{X}} = \mu \quad (13.32)$$

$$\sigma_{\bar{X}}^2 = \begin{cases} \frac{\sigma^2}{n}; & \text{for samples from an infinite population} \\ \frac{\sigma^2}{n} \left( \frac{N-n}{N-1} \right); & \text{for samples from a population of size } N \end{cases} \quad (13.33)$$

These results are straightforward to establish.

For infinite populations, the standard deviation of the sample mean,  $\sigma_{\bar{X}}$ , is:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \quad (13.34)$$

known as the *standard error of the mean*.

### 13.3.3 Limiting Distribution of the Mean

As shown in the last subsection, if the underlying population pdf is unknown, the full sampling distribution of  $\bar{X}$ , the mean of a random sample drawn from this population, cannot be determined; but the mean and variance of the sampling distribution are known. Nevertheless, even though the complete sampling distribution is unknown in general, we know the limiting distribution (as  $n \rightarrow \infty$ ) of a closely related random variable, the standardized mean defined as:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad (13.35)$$

The limiting distribution of  $Z$  is given by the following theorem.

**The Central Limit Theorem (CLT):** Let  $\bar{X}$  be the mean of the random sample  $(X_1, X_2, \dots, X_n)$  taken from a population with mean,  $\mu$ , and (finite) variance  $\sigma^2$ . Define the random variable  $Z$  according to Eq (13.35); then the pdf of  $Z$ ,  $f(z)$ , tends to  $N(0, 1)$ , a standard normal distribution, in the limit as  $n \rightarrow \infty$ .

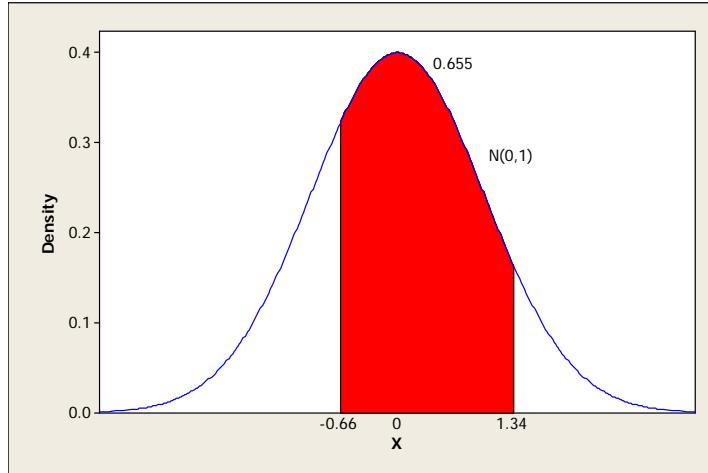
**Remarks:**

1. Regardless of the distribution underlying the original population from which the random sample was drawn, the distribution of the sample mean approaches a normal distribution as  $n \rightarrow \infty$ . In fact, for  $n$  as small as 25 or 30, the normal approximation can be quite good.
2. The random variable  $\sqrt{n}(\bar{X} - \mu)/\sigma$  is approximately distributed  $N(0, 1)$ ; it will therefore be possible to employ the standard normal distribution,  $N(0, 1)$ , to obtain approximate probabilities concerning  $\bar{X}$ .
3. If the original population has a  $N(\mu, \sigma^2)$  distribution, then, it can be shown that the random variable  $Z$ , the standardized mean, defined in Eq (13.35), has *exactly* the  $N(0, 1)$  distribution.

We are now in a position to consider how this result might find application in statistical inference about the mean  $\mu$  of a population. Consider a random sample  $(X_1, X_2, \dots, X_n)$ , from which the sample mean,  $\bar{X}$ , is computed: if the population variance is known, then regardless of the underlying population pdf, the standard normal distribution can be used to determine probabilities about  $\bar{X}$  indirectly via the standardized mean. Let us illustrate this point with the following example.

**Example 13.3: PROBABILITIES CONCERNING MEAN LIFETIMES OF DIGITAL LIGHT PROCESSING (DLP) PROJECTOR LAMPS**

As justification for its high price, an expensive brand of DLP projector lamps has been purported to have an impressively long average lifetime of 5,133 hrs, with a standard deviation of 315 hrs. (1) In preparing to send a trial sample of 40 to a customer, the company manufacturing the lamps wishes to accompany the shipment with a “factory specification” stating the probability that the mean lifetime for this sample will be between the round numbers 5,100 and 5,200 hrs. Determine this probability. (2) In a follow-up to the shipment, information from end-use purchasers was used to track the actual lifetimes of the 40 LCD projector lamps; the result showed an average lifetime less than 5,000 hours. Compute the probability of obtaining such a low average lifetime



**FIGURE 13.1:** Sampling distribution for mean lifetime of DLP lamps in Example 13.3 used to compute  $P(5100 < \bar{X} < 5200) = P(-0.66 < Z < 1.34)$

by chance alone, if the lamps truly came from a collection (population) with  $\mu = 5,133$  and  $\sigma = 315$ .

**Solution:**

(1) The problem requires that we determine  $P(5100 < \bar{X} < 5200)$ , but with no specified pdf, we cannot calculate this probability directly. Nevertheless, knowing that the standardized mean,  $Z$ , has a  $N(0, 1)$  distribution allows us to compute the approximate probability as follows:

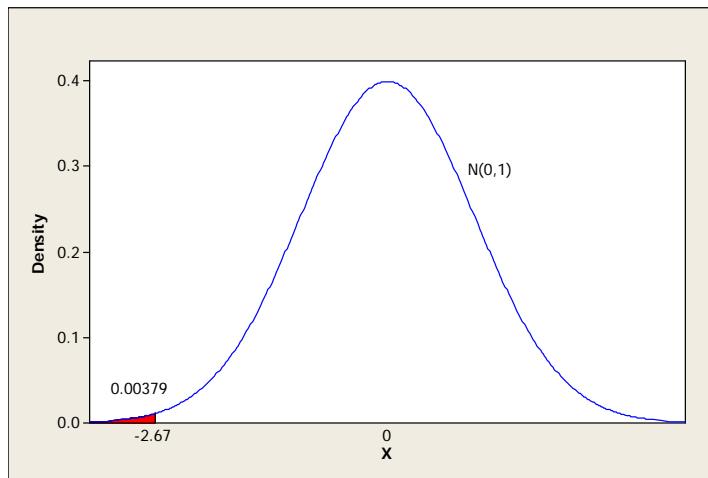
$$\begin{aligned} P(5100 < \bar{X} < 5200) &= P\left\{\left(\frac{5100 - 5133}{315/\sqrt{40}}\right) < Z < \left(\frac{5200 - 5133}{315/\sqrt{40}}\right)\right\} \\ &= P(-0.66 < Z < 1.34) = 0.655 \end{aligned} \quad (13.36)$$

where the indicated probability has been computed directly from the computer program MINITAB using the cumulative probability calculation option for a Normal distribution (**Calc > Prob Dist > Normal**), with mean = 0, standard deviation = 1, to obtain  $F_Z(1.34) = 0.910$  and  $F_Z(-0.66) = 0.255$  yielding the indicated result. (See Fig 13.1). Tables of standard normal probabilities could also be used to obtain this result.

Thus there is a 65.5% chance that the actual average lifetime will be between 5,100 and 5,200 hours if the lamps truly came from a population with  $\mu = 5,133$  and  $\sigma = 315$ .

(2) Employing the same approximate  $N(0, 1)$  distribution for the standardized mean, we are able to compute the required  $P(\bar{X} < 5000)$  as follows:

$$\begin{aligned} P(\bar{X} < 5000) &= P\left\{\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} < \frac{\sqrt{40}(5000 - 5133)}{315}\right\} \\ &= P(Z < -2.67) = 0.004 \end{aligned} \quad (13.37)$$



**FIGURE 13.2:** Sampling distribution for average lifetime of DLP lamps in Example 13.3 used to compute  $P(\bar{X} < 5000) = P(Z < -2.67)$

where the probability is obtained directly from MINITAB again using the cumulative probability calculation option for a Normal distribution (**Calc > Prob Dist > Normal**), with mean = 0, standard deviation = 1, to obtain  $F_Z(-2.67) = 0.004$  (see Fig 13.2).

And now, because the probability of obtaining — by chance alone — a sample of 40 lamps with such a low average lifetime (< 5,000 hours) from a population purported to have a much higher average lifetime, is so small, it is very doubtful that this sample came from the postulated population. It appears more likely that the result from this sample is more representative of the true lifetimes of the lamps. If true, then the practical implication is that there is reason to doubt the claim that these DLP projector lamps truly merit the purported “long lifetime” characterization.

This example is a preview of what is ahead in Chapters 14 and 15, hinting at some of the principles used in formal statistical inference.

### 13.3.4 $\sigma$ Unknown

When the underlying population distribution is unknown, the concept of employing a limiting distribution for  $\bar{X}$  as discussed in the previous subsection, works only if the population variance is known; only then can the standardized mean,  $Z$ , defined as in Eq (13.35) be a bona-fide “statistic.” If  $\sigma^2$  is unknown,  $Z$  is no longer a proper “statistic” because it will then contain an unknown population parameter. (It is important to realize that in the current context, the sampling distribution of  $\bar{X}$  is with respect to a *postulated* population

mean,  $\mu$ , that  $\bar{X}$  is supposed to represent; with  $\mu$  specified, it is no longer an unknown population parameter.)

When  $\sigma^2$  is unknown, we have two options: (i) if the sample size,  $n$ , is large (say  $n \geq 50$ ),  $S^2$ , the sample variance, provides a good approximation to  $\sigma^2$ ; (ii) when  $n$  is small, it seems reasonable to contemplate using the random variable,  $\sqrt{n}(\bar{X} - \mu)/S$ , which is the standardized mean,  $Z$ , with the unknown population standard deviation replaced with the sample standard deviation,  $S$ , defined as:

$$S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}} \quad (13.38)$$

Unfortunately, nothing can be said *in general* about the distribution of this random variable if the underlying population distribution is unknown, and  $n$  is small. However, if the sample comes from a population having a Gaussian distribution (the so-called normal population), then the following result holds:

Let  $\bar{X}$  and  $S$  be, respectively, the mean and standard deviation of the random sample  $(X_1, X_2, \dots, X_n)$  of size  $n$  drawn from a population with distribution  $N(\mu, \sigma^2)$ ; then the statistic

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \quad (13.39)$$

has a Student's  $t$ -distribution with  $\nu = n - 1$  degrees of freedom.

### Remarks:

1. We encountered this result in Chapter 9 (section 9.3.5) during our discussion of probability models for continuous random variables.
2. This result is somewhat more general than the Central Limit Theorem because it does not require knowledge of  $\sigma^2$ ; conversely, it is less general in that it requires the “normality assumption” for the underlying distribution, which the CLT does not require.
3. The result holds *exactly* for any  $n$ : under the normality assumption, the pdf of  $T$  is exactly the  $t$ -distribution regardless of sample size  $n$ ; the CLT on the other hand prescribes a limiting distribution as  $n \rightarrow \infty$ .
4. As noted earlier, the  $t$ -distribution approaches the standard normal distribution as  $\nu$  (hence,  $n \rightarrow \infty$ ).
5. The “normality assumption” is not too severe, however; when samples

are from non-normal populations, the distribution of the  $T$  statistic is still fairly close to the Student's  $t$ -distribution.

Let us illustrate the application of this result with the following example.

**Example 13.4: MEAN DIAMETER OF BALL BEARINGS**

A manufacturer of "low precision" 10 mm diameter ball bearings periodically takes samples and measures them to confirm that the manufacturing process is still on target. A random sample of 20 ball bearings resulted in diameter measurements with an average of  $\bar{x} = 9.86$  mm, and a standard deviation  $s = 1.01$  mm. Postulating that the random sample  $X_1, X_2, \dots, X_{20}$  is from a Gaussian distribution with  $\mu = 10$ , find the probability that any sample mean,  $\bar{X}$ , will fall to either side of the postulated mean by the observed amount or more, by chance alone. i.e.  $P(\bar{X} \leq 9.86)$  or  $P(\bar{X} \geq 10.14)$ .

**Solution:**

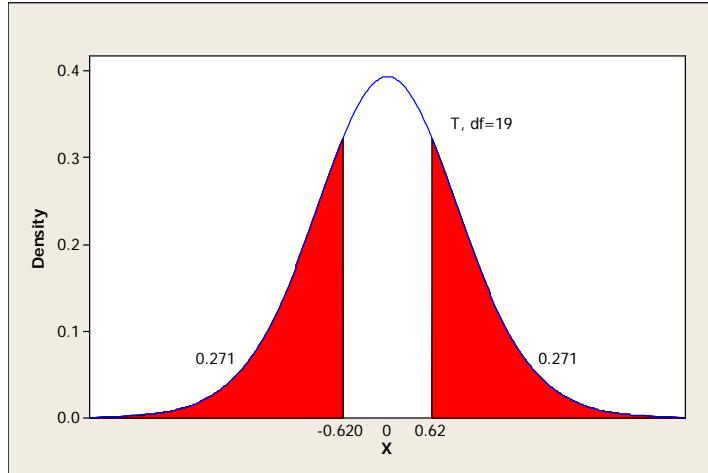
Had the population variance,  $\sigma^2$ , been given, we could have obtained the sampling distribution for  $\bar{X}$  precisely (as another normal distribution with  $\mu = 10$  and variance  $\sigma^2/20$ ); these results could have been used directly to compute the required probabilities. However, since the population variance is not given, the required probability  $P(\bar{X} \leq 9.86) + P(\bar{X} \geq 10.14)$  is determined using the  $T$  statistic:

$$\begin{aligned} P(\bar{X} \leq 9.86) + P(\bar{X} \geq 10.14) &= P\left\{\frac{\sqrt{n}|\bar{X} - \mu|}{S} \leq \frac{\sqrt{20}|9.86 - 10|}{1.01}\right\} \\ &= P(T \leq -0.62) + P(T \geq 0.62) \\ &= 0.542 \end{aligned} \quad (13.40)$$

Again, the probabilities are obtained directly from MINITAB using the cumulative probability computation option for the  $t$ -distribution, with  $\nu = 19$  degrees of freedom to give  $F_T(-0.62) = 0.271$ , and by symmetry,  $P(T \geq 0.62) = 0.271$  to obtain the result in Eq (13.40). (See Fig 13.3.)

The implication is that, under the postulate that the ball bearings are truly 10 mm in diameter, there is a fairly substantial 54% chance that the observed sample average misses the target of 10 mm to the left (comes in as 9.86 or less) or to the right by the same amount (comes in as 10.14 or more) *purely at random*. In other words, by pure chance alone, one would expect to see this sort of observed deviation of the sample mean from the true postulated (target) mean diameter value of 10 mm, more than half the time. The inclination therefore is to conclude that there is *no evidence* in this sample data that the process is off-target.

Again, as with the previous example, this one also provides a preview of what is ahead in Chapters 14 and 15.



**FIGURE 13.3:** Sampling distribution of the mean diameter of ball bearings in Example 13.4 used to compute  $P(|\bar{X} - 10| \geq 0.14) = P(|T| \geq 0.62)$

### 13.4 Sampling Distribution of the Variance

Similar to the preceding discussion on the distribution of the mean,  $\bar{X}$ , of a random sample  $(X_1, X_2, \dots, X_n)$ , we observe that the sample variance, defined as,

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{(n-1)} \quad (13.41)$$

is a random variable that is clearly related to the population variance, but whose specific value,  $s^2$ , will vary from sample to sample. To make objective inference statements about the population variance from this sample quantity, we also need a theoretical distribution of  $S^2$ . For this we have the following three results:

1. **Sample Variance:** Let  $X_1, X_2, \dots, X_n$  be a random sample taken from a population with an arbitrary pdf,  $f(x)$ , having a mean  $\mu$  and variance  $\sigma^2$ . The sample variance defined as in Eq (13.41) is itself a random variable whose expectation is  $\sigma^2$ ; i.e.

$$E[S^2] = \sigma^2 \quad (13.42)$$

2. **Sampling Distribution of Single Variance:** If the random sample  $(X_1, X_2, \dots, X_n)$  is from a normal population, then the random variable:

$$C = \frac{(n-1)S^2}{\sigma^2} \quad (13.43)$$

has a  $\chi^2(n - 1)$  distribution. Such a distribution can be used to compute probabilities concerning the random variable,  $S^2$ . Unfortunately, nothing so explicit can be said about sampling distributions of variances for random samples drawn from non-normal populations.

3. **Sampling Distribution of Two Variances:** Let  $S_1^2$  and  $S_2^2$  be the variances of two sets of independent random samples of respective sizes  $n_1$  and  $n_2$ , each drawn from two normal populations *having the same variance*. Then the random variable:

$$F = \frac{S_1^2}{S_2^2} \quad (13.44)$$

has the Fisher  $F(\nu_1, \nu_2)$  distribution, with degrees of freedom  $\nu_1 = (n_1 - 1)$  and  $\nu_2 = (n_2 - 1)$ . Again, for samples from non-normal distributions, these results do not hold.

As with the sampling distributions of the mean, it has been customary to compute probabilities from these distributions using  $\chi^2(n - 1)$  and  $F$  tables. It is now more common to use computers instead of tables, as illustrated with the following examples.

#### **Example 13.5: VARIANCE OF BALL BEARINGS DIAMETER MEASUREMENTS**

The random sample of 20 ball bearings diameters  $X_1, X_2, \dots, X_{20}$ , postulated in Example 13.4 as being from a Gaussian distribution with  $\mu = 10$ , resulted in an average measured diameter of  $\bar{x} = 9.86$  mm, and a standard deviation  $s = 1.01$  mm. Now consider the case in which the design standard deviation for the manufacturing process is specified as  $\sigma = 0.9$ ; compute the probability that any sample standard deviation will equal or exceed the observed value of  $S = 1.01$  mm if the manufacturing process is still operating true to the original design specifications.

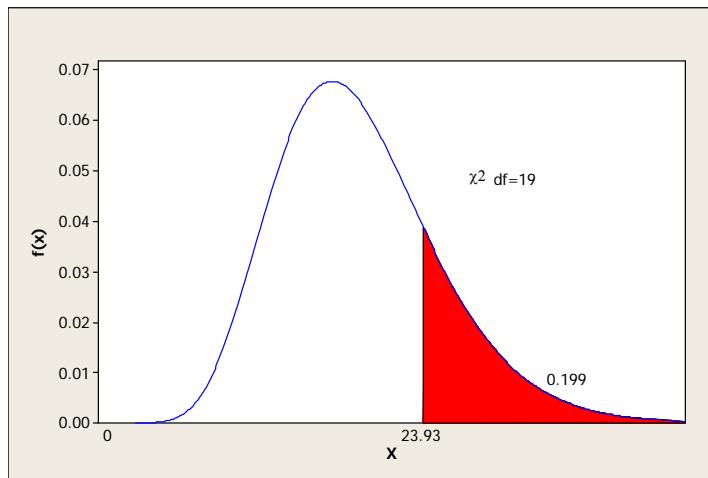
#### **Solution:**

This problem requires computing the probability  $P(S \geq 1.01)$  using the chi-square statistic,  $C$ ; i.e.

$$\begin{aligned} P(S \geq 1.01) &= P\left(C \geq \frac{19(1.01)^2}{(0.9)^2}\right) \\ &= P(C \geq 23.93) = 0.199 \end{aligned} \quad (13.45)$$

where the indicated probability was obtained directly from MINITAB as with all the other earlier examples: the cumulative probability calculation option for a Chi-square distribution (**Calc > Prob Dist > Chi-Square**), with degrees of freedom = 19; and “input constant” 23.93 to obtain  $P(C \leq 23.93) = 0.801$  to yield  $P(C \geq 23.93) = 1 - 0.801 = 0.199$ . (See Fig 13.4.)

Thus, under the postulate that the design standard deviation is 0.9, the sampling distribution of the variance indicates a fairly high 20%



**FIGURE 13.4:** Sampling distribution for the variance of ball bearing diameters in Example 13.5 used to compute  $P(S \geq 1.01) = P(C \geq 23.93)$

chance of obtaining, purely at random, sample variances that are 1.01 or higher, even when the process is operating as designed.

**Example 13.6: VARIANCE OF BALL BEARINGS DIAMETER MEASUREMENTS: TWO RANDOM SAMPLES**

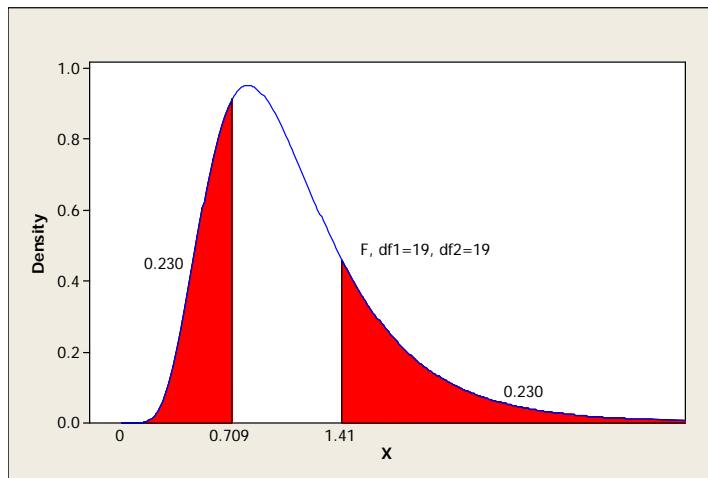
A second random sample of 20 ball bearings taken a month after the sample examined in Examples 13.3 and 13.4 yielded an average measured diameter of  $\bar{x}_2 = 10.03$  mm, and a standard deviation  $s_2 = 0.85$  mm. Find the probability that the process operation remains essentially unchanged, in terms of the observed sample standard deviation, even though the newly observed sample standard deviation is less than the value observed a month earlier. All assumptions in Example 13.4 hold.

**Solution:**

In this case, we are concerned with comparing *two* sample variances,  $S_1^2$  from the previous month (with the specific value of  $(1.01)^2$ ) and  $S_2^2$ , the most recent, with specific value  $(0.85)^2$ . Since the real question is not whether one value is greater than the other, but whether the two sample variances *are equal or not*, we use the *F*-distribution with degrees of freedom (19,19) to obtain the probability that  $F \geq (1.01)^2/(0.85)^2$  (*or, vice-versa, that  $F \leq (0.85)^2/(1.01)^2$* ). The required probability is obtained as:

$$P(F \geq 1.41) + P(F \leq 0.709) = 0.460$$

The indicated probabilities are, once again, obtained directly from MINITAB for the *F*-distribution with  $\nu_1 = \nu_2 = 19$  degrees of freedom; (See Fig 13.5.) The implication is that there is almost a 50% chance that the observed difference between the two sample variances



**FIGURE 13.5:** Sampling distribution for the two variances of ball bearing diameters in Example 13.6 used to compute  $P(F \geq 1.41) + P(F \leq 0.709)$

will occur purely at random. We conclude, therefore, that there does not appear to be any evidence that the process operation has changed in the past month since the last sample was taken.

With these concepts in place, we are now in a position to discuss the two aspects of statistical inference, beginning with Estimation in the next chapter, followed by Hypothesis Testing in Chapter 15.

### 13.5 Summary and Conclusions

Because of its foundational role in inductive statistics, sampling—the study and characterization of samples drawn from a population—was considered in this chapter by itself first, as a precursor to a formal treatment of inductive statistics *proper*, i.e., estimation and hypothesis testing. The entire chapter was devoted to one problem: how to obtain a probability distribution for the sample (or more precisely, functions thereof), given the population distribution.

The concept of the “statistic” is central to sampling, although the full import will not be realized until the next chapter when we consider the important problem of how to determine unknown population parameters from sample data. For the purpose of this chapter, it was sufficient to introduce the “statistic” simply as any function of a random sample that does not contain unknown population parameters; the rest of the chapter was then devoted to

determining the distribution of such “statistics.” In particular, since in inductive statistics (as we shall soon discover in Chapters 14 and 15), the two most important “statistics” are the sample mean and sample variance, the distributions of these quantities were characterized under various conditions, giving rise to results—some general, others specific only to normal populations—that are used extensively in the next two chapters. Of all the general results, the one used most frequently arises from the Central Limit Theorem, through which we know that, regardless of the underlying distribution, as the sample size tends to infinity, the sampling distribution of the sample mean tends to the Gaussian distribution. This collection of results provides the foundation for the next two chapters.

Here are some of the main points of the chapter again.

- Sampling is concerned with the probabilistic characterization of finite size samples drawn from a population; it is the statistical analog to the probability problem of characterizing individual observations of a random variable with a pdf.
- The central concepts in sampling are:
  - *The random sample:* in principle,  $n$  independent random variables drawn from a population in such a way that each one has an equal chance of being drawn; the mathematical consequence is that if  $f(x_i)$  is the pdf for the  $i^{th}$  random variable, then the joint pdf of the random sample is a product of the contributing pdfs;
  - *The “statistic:”* a function of one or more random variables that does not contain an unknown population parameter.
  - *The sampling distribution:* the probability distribution of a statistic of interest; its determination is significantly facilitated if the statistic is a function of a *random sample*.
- As a consequence of the central limit theorem, we have the general result that as  $n \rightarrow \infty$ , the distribution of the mean of a random sample drawn from any population with known mean and variance, is Gaussian, greatly enabling the probabilistic analysis of means of large samples.

---

## REVIEW QUESTIONS

1. What is sampling?
2. As presented in Section 13.1 what are the three central concepts of sampling theory?

- 3.** What is a random sample? And what is the mathematical implication of the statement that  $X_1, X_2, \dots, X_n$  constitutes a random sample from a distribution that has a pdf  $f(x)$ ?
- 4.** What is a “statistic”?
- 5.** What is a sampling distribution?
- 6.** What is the primary utility of a statistic and its distribution?
- 7.** How is the discussion in Chapter 6 helpful in determining sampling distributions?
- 8.** What is the sampling distribution of a linear combination of  $n$  independent Gaussian random variables with identical pdfs?
- 9.** What is the sampling distribution of a sum of  $n$  independent  $\chi^2(r)$  random variables with identical pdfs?
- 10.** If  $\bar{X}$  is the mean of a random sample of size  $n$  from a population with mean  $\mu_X$  and variance  $\sigma_X^2$ , what is  $E(\bar{X})$  and what is  $Var(\bar{X})$ ?
- 11.** What is the standard error of the mean?
- 12.** What is the central limit theorem as stated in Section 13.3? What are its implications in sampling theory?
- 13.** In sampling theory, under what circumstances will the  $t$ -distribution be used instead of the standard normal distribution?
- 14.** What is the sampling distribution of the variance of a random sample of size  $n$  drawn from a Gaussian distribution?
- 15.** What is the sampling distribution of the ratio of the variances of two sets of independent random samples of sizes  $n_1$  and  $n_2$  each drawn from two normal populations having the same variance?

## EXERCISES

### Section 13.1

**13.1** Given  $X_1, X_2, \dots, X_n$ , a random sample from a population with mean  $\mu$  and variance  $\sigma^2$ , both unknown, determine which of the following functions of this random sample is a statistic and which is not.

- (i)  $Y_1 = (\prod_{i=1}^n X_i)^{1/n}$ ;
- (ii)  $Y_2 = \sum_{i=1}^n (X_i - \mu)^2$ ;
- (iii)  $Y_3 = \sum_{i=1}^n \omega_i X_i$ ;  $\sum_{i=1}^n \omega_i = 1$ ;

$$(iv) Y_4 = \sum_{i=1}^n \frac{X_i}{\sigma}.$$

If the population mean  $\mu$  is specified, how will this change your answer?

**13.2** Given  $X_1, X_2, \dots, X_n$ , a random sample from a population with mean  $\mu$  and variance  $\sigma^2$ , both unknown, define the following statistic:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

as the sample mean. Determine which of the following functions of the random variable are statistics and which are not:

- (i)  $Y_1 = \sum_{i=1}^n (X_i - \bar{X})^2 / (n - 1)$ ;
- (ii)  $Y_2 = \sum_{i=1}^n (X_i - \mu)^2 / n$ ;
- (iii)  $Y_3 = \sum_{i=1}^n |X_i - \bar{X}|^k / (n - 1)$ ;  $\forall k > 0$ ;
- (iv)  $Y_4 = \sum_{i=1}^n (X_i - \bar{X}) / \sigma$ .

**13.3** For each of the following distributions, given the population mean  $\mu$ , and variance  $\sigma^2$ , derive the appropriate expressions for obtaining the actual pdf parameters ( $\alpha, \beta$ , or  $n, p$ ) in terms of  $\mu$  and  $\sigma^2$ : (i) Gamma( $\alpha, \beta$ ); (ii) Beta( $\alpha, \beta$ ); (iii) Binomial( $n, p$ ).

### Section 13.2

**13.4** Given  $n$  mutually stochastically independent random variables,  $X_1, X_2, \dots, X_n$ , with respective pdfs  $N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2), \dots, N(\mu_n, \sigma_n^2)$ :

- (i) Determine the distribution of the statistic:

$$Y = k_1 X_1 + k_2 X_2 + \dots + k_n X_n$$

where  $k_1, k_2, \dots, k_n$  are real constants; and show that it is a Gaussian distribution,  $N(\mu_y, \sigma_y^2)$  where

$$\begin{aligned}\mu_y &= k_1 \mu_1 + k_2 \mu_2 + \dots + k_n \mu_n \\ \sigma_y^2 &= k_1^2 \sigma_1^2 + k_2^2 \sigma_2^2 + \dots + k_n^2 \sigma_n^2\end{aligned}$$

(ii) From this result, show that if  $\mu_1 = \mu_2 = \dots = \mu_n = \mu$ , and  $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_n^2 = \sigma^2$ , so that the random variables  $X_i; i = 1, 2, \dots, n$  are all identically distributed, then establish the result given in Eqs (13.15) and (13.16), i.e., that

$$\begin{aligned}\mu_y &= \left( \sum_{i=1}^n k_i \right) \mu \\ \sigma_y^2 &= \left( \sum_{i=1}^n k_i^2 \right) \sigma^2\end{aligned}$$

**13.5** Given  $n$  mutually stochastically independent random variables,  $X_1, X_2, \dots, X_n$ , with identical pdfs that are unspecified except that the mean is  $\mu$ , and the variance,  $\sigma^2$ ; show that the mean and variance of the statistic defined as

$$Y = \frac{1}{n} \sum_{i=1}^n X_i$$

are given by

$$\begin{aligned}\mu_y &= \mu \\ \sigma_y^2 &= \sigma^2/n\end{aligned}$$

and hence establish the results given in Eqs (13.17) and (13.18) in the text.

**13.6** Given a random sample  $X_1, X_2, \dots, X_n$ , from a Gaussian  $N(\mu, \sigma^2)$  distribution; show that the random variable

$$Y = \sum_{i=1}^n \left( \frac{X_i - \mu}{\sigma} \right)^2$$

has a  $\chi^2(n)$  distribution.

**13.7** Given  $n$  mutually stochastically independent random variables,  $X_1, X_2, \dots, X_n$ , with respective pdfs  $\chi^2(r_1), \chi^2(r_2), \dots, \chi^2(r_n)$ ; show that the random variable

$$Y = X_1 + X_2 + \dots + X_n$$

has a  $\chi^2(r)$  distribution with degrees of freedom,

$$r = r_1 + r_2 + \dots + r_n.$$

**13.8** Given a random sample  $X_1, X_2, \dots, X_n$  from a Poisson  $\mathcal{P}(\lambda)$  distribution, determine the sampling distribution of the random variable defined as

$$Y = \frac{1}{n} \sum_{i=1}^n X_i$$

**13.9** Given a random sample  $X_1, X_2, \dots, X_n$  from a Gamma  $\gamma(\alpha, \beta)$  distribution, determine the sampling distribution of the random variable defined as

$$Y = \sum_{i=1}^n X_i$$

### Section 13.3

**13.10** Given that  $X_1, X_2, \dots, X_n$  constitute a random sample from a population with mean  $\mu$ , and variance  $\sigma^2$ , define two statistics representing the sample mean as follows:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (13.46)$$

$$\tilde{X} = \sum_{i=1}^n \omega_i X_i; \text{ with } \sum_{i=1}^n \omega_i = 1 \quad (13.47)$$

where the first is a regular mean and the second is a “weighted” mean. Show that  $E(\bar{X}) = \mu$  and also that  $E(\tilde{X}) = \mu$ ; but that  $Var(\bar{X}) \leq Var(\tilde{X})$ .

**13.11** If  $X_1, X_2, \dots, X_n$  is a random sample from a Poisson  $\mathcal{P}(\lambda)$  distribution, find the distribution of  $\bar{X}$ , the sample mean defined in Eq 14.69. (Hint: See Example 6.1

in Chapter 6.) Determine  $E(\bar{X})$  and  $Var(\bar{X})$ .

**13.12** If  $X_1, X_2, \dots, X_n$  is a random sample from a Gamma  $\gamma(\alpha, \beta)$  distribution, find the distribution of  $\bar{X}$  the sample mean defined in Eq 14.69. Determine  $E(\bar{X})$  and  $Var(\bar{X})$ .

**13.13** If  $X_1, X_2, \dots, X_n$  is a random sample from a Lognormal  $\mathcal{L}(\alpha, \beta)$  distribution,

- (i) Find the distribution of the geometric mean

$$\bar{X}_g = \left( \prod_{i=1}^n X_i \right)^{1/n} \quad (13.48)$$

(ii) Determine  $E(\ln \bar{X}_g)$  and  $Var(\ln \bar{X}_g)$ .

**13.14** Given a random sample of 10 observations drawn from a Gaussian population with mean 100, and variance 25, compute the following probabilities regarding the sample mean,  $\bar{X}$ :

- (i)  $P(\bar{X} \geq 100)$ ; (ii)  $P(\bar{X} \leq 100)$ ; (iii)  $P(\bar{X} \geq 104.5)$ ; (iv)  $P(96.9 \leq \bar{X} \leq 103.1)$ ; (v)  $P(98.4 \leq \bar{X} \leq 101.6)$ . Will the sample size make a difference in the distribution used to compute the probabilities?

**13.15** Refer to Exercise 13.14. This time, the population variance is not given; instead, the sample variance was obtained as 24.7 from the 10 observations.

- (i) Compute the five probabilities.
- (ii) Recompute the probabilities if the sample size increased to 30 but the sample variance remained the same.

**13.16** A sample of size  $n$  is drawn from a large population with mean  $\mu$  and variance  $\sigma^2$  but unknown distribution;

- (i) Determine the mean and variance of the sample mean when  $n = 10; \mu = 50; \sigma^2 = 20$ ;
- (ii) Determine the mean and variance of the sample mean when  $n = 20; \mu = 50; \sigma^2 = 20$ ;
- (iii) Determine the probability that a sample mean obtained from a sample of size  $n = 50$  will *not* deviate from the population mean by more than  $\pm 3$ . State any assumption you may need to make in answering this question.

**13.17** A random sample of size  $n = 50$  is taken from a large population with mean 15 and variance 4, but unknown distribution.

- (i) What is the standard deviation  $\sigma_{\bar{X}}$  of the sample mean?
- (ii) If the sample size were reduced by 50%, what will be the new standard deviation  $\sigma_{\bar{X}}$  of the sample mean?
- (iii) To reduce the standard deviation to 50% of the value in (i), what sample size will be needed?

### Section 13.4

**13.18** The variance of a sample of size  $n = 20$  drawn from a normal population with mean 100 and variance 10 was obtained as  $s^2 = 9.5$ .

- (i) Determine the probability that  $S^2 \leq 10$ .

- (ii) Determine the probability that  $S^2 \leq 9.5$ .
- (iii) If the sample size increased by 50% but the computed sample variance remained the same, recompute the probability  $S^2 \leq 9.5$ .
- (iv) Repeat part (iii) when the sample size is decreased from the original value of 50 by 50%.

**13.19** Two samples were taken from the same normal population with mean 100 and variance 10. One sample, of size 20, had a sample variance  $S_1^2 = 11.2$ ; the other of size 30, a sample variance of  $S_2^2 = 9.8$ .

- (i) Determine the following probabilities:  $P(S_1^2 \leq 9.8)$  and  $P(S_2^2 \geq 11.2)$ .
- (ii) Determine the specific variate  $\chi_0^2$  for this problem such that for  $n = 20$ ,  $P(C \geq \chi_0^2) = 0.05$  where  $C$  is a  $\chi^2(n)$  distributed random variable.
- (iii) Determine the specific variate  $\chi_0^2$  for this problem such that for  $n = 30$ ,  $P(C \leq \chi_0^2) = 0.05$  where, as in (ii),  $C$  is a  $\chi^2(n)$  distributed random variable.

**13.20** Refer to Exercise 13.19. Use an appropriate distribution to determine the probability of observing a ratio  $11.2/9.8$ , or greater, if the two sample variances are equal.

**13.21** Two samples of equal size  $n = 51$  are drawn from normal populations with the same variance. One sample variance was  $S_1^2 = 15$ ; the other,  $S_2^2 = 12.0$ . Determine the following probabilities:

- (i)  $P(S_1^2/S_2^2 \geq 1)$ ; and  $P(S_2^2/S_1^2 \leq 1)$
- (ii)  $P(S_1^2/S_2^2 \geq 1.25)$  and  $P(S_2^2/S_1^2 \leq 0.8)$

## APPLICATION PROBLEMS

**13.22** The following data set, from a study by Lucas (1985)<sup>1</sup>, shows the number of accidents occurring per quarter (three months) at a DuPont company facility, over a 10-year period. The data set has been partitioned into two periods: Period I is the first five-year period of the study; Period II, the second five-year period.

| Period I |   |    |    | Period II |   |   |   |
|----------|---|----|----|-----------|---|---|---|
| 5        | 5 | 10 | 8  | 3         | 4 | 2 | 0 |
| 4        | 5 | 7  | 3  | 1         | 3 | 2 | 2 |
| 2        | 8 | 6  | 9  | 7         | 7 | 1 | 4 |
| 5        | 6 | 5  | 10 | 1         | 2 | 2 | 1 |
| 6        | 3 | 3  | 10 | 4         | 4 | 4 | 4 |

- (i) Why is a Poisson pdf a reasonable model for representing this data?
- (ii) Treat the entire 10-year data as a single block and the data shown as specific observations  $\{x_i\}_{i=1}^{40}$  of a random sample  $X_i; i = 1, 2, \dots, 40$ , from a single Poisson population with unknown parameter  $\lambda$ . Obtain a numerical value for the sample mean,  $\bar{X}$ ; use it as an estimate of the unknown population parameter,  $\lambda$ , to produce

<sup>1</sup>Lucas J. M., (1985). "Counted Data CUSUMs," *Technometrics*, 27, 129–144

a complete pdf,  $f(x|\lambda)$ , as a candidate description of the population. From this complete pdf, compute the probabilities  $f(x)$  for  $x = 1, 2, \dots, 10$ .

- (iii) Using the value estimated for the population parameter in (ii), determine the precise (not approximate) distribution of  $\bar{X}$ . (See Exercise 13.11.) From this distribution for  $\bar{X}$ , compute  $P(\bar{X} \geq 3)$  and  $P(\bar{X} \leq 6)$ .
- (iv) Using the candidate population description obtained in (ii), compute  $P(X \leq 6)$  and  $P(X \geq 3)$ . (Note that these probabilities are with respect to the random variable  $X$  itself, not the sample mean.) Are the 20 observations in Period I and those in Period II consistent with these theoretical probabilities? Comment on what these results imply about whether or not the postulated population model is plausible.

**13.23** Refer to Application Problem 13.22. This time consider each period as two separate blocks of specific observations  $\{x_i\}_{i=1}^{20}$  of a random sample  $X_i; i = 1, 2, \dots, 20$  from two distinct Poisson populations with unknown parameters  $\lambda_1$  for Period I and  $\lambda_2$  for Period II. Obtain numerical values for the two sample means,  $\bar{X}_1$  and  $\bar{X}_2$  and use these as estimates  $\hat{\lambda}_1$  and  $\hat{\lambda}_2$  for population parameters  $\lambda_1$  and  $\lambda_2$  respectively. Determine the exact sampling distributions for  $\bar{X}_1$  and  $\bar{X}_2$  and use these to compute the following probabilities:

- (i)  $P(\bar{X}_1 \geq 3|\lambda_1 = \hat{\lambda}_1)$  and
- (ii)  $P(\bar{X}_2 \leq 6|\lambda_2 = \hat{\lambda}_2)$ .

Comment on what these results mean in terms of the conjecture that these two populations may in fact be different.

**13.24** Refer to Application Problem 13.22 and consider the two propositions that (i) the data for Period I represent DuPont facility operation when the true mean number of accident occurrences per quarter is 6; and that (ii) by Period II, the true mean number of accident occurrences has been reduced by 50%. By computing appropriate probabilities and carrying out appropriate comparisons, argue for or against the statement: “*These two propositions are consistent with the data.*”

**13.25** Refer to Application Problem 13.22 and consider the data from the two periods separately as random samples from two distinct Poisson populations with parameters  $\lambda = 6$  for Period I and  $\lambda = 3$  for Period II

- (i) Using the exact sampling distribution for  $\bar{X}_1$  for Period I and  $\bar{X}_2$  for Period II, determine  $P(\bar{X}_1 \geq 4.5)$  and  $P(\bar{X}_2 \geq 4.5)$
- (ii) Use the approximate Gaussian result arising from the central limit theorem to obtain the sampling distributions for  $\bar{X}_1$  for Period I and  $\bar{X}_2$  for Period II; recompute the probabilities in (i) and compare the two sets of results. Comment on the accuracy of the approximate Gaussian sampling distributions in this particular case.

**13.26** The waiting time (in days) until the occurrence of a recordable safety incident in a certain company’s manufacturing site is known to be an exponential random variable with unknown parameter  $\beta$ . As part of a safety performance characterization program, the following data set was obtained

$$S_1 = \{16, 1, 9, 34, 63, 44, 1, 63, 41, 29\}$$

- (i) Considering this as a specific realization of a random sample of size  $n = 10$ , determine a numerical value,  $\bar{x}$ , for the random sample mean,  $\bar{X}$ , and obtain an exact

sampling distribution for the sampling mean in terms of the unknown population parameter,  $\beta$ , and sample size,  $n$ .

- (ii) The company claims that its true mean waiting time between safety incidents is 40 days. From the sampling distribution and the specific value  $\bar{x}$  obtained in (i), determine  $P(\bar{X} \leq \bar{x}|\beta = 40)$ .
- (iii) An independent safety auditor estimates that from the company's records, operating procedures, and other performance characteristics, the true mean waiting time between safety incidents is more likely to be 30 days. Determine  $P(\bar{X} \leq \bar{x}|\beta = 30)$ . Compare this probability to the one obtained in (ii) and comment on which postulated value for the true mean waiting time is more believable,  $\beta = 40$  as claimed by the company, or  $\beta = 30$  as claimed by the independent safety auditor.

**13.27** Refer to Application Problem 13.26. From the pdf of the exponential  $\mathcal{E}(40)$  random variable, determine  $P(X \leq 10)$ . Recompute  $P(X \leq 10)$  using the pdf for the exponential  $\mathcal{E}(30)$  random variable. Which of these results is more consistent with the data set  $S_1$ ? Comment on what this implies about the more likely value for the population parameter.

**13.28** Refer to the data in Table 1.1 in Chapter 1, which shows 50 samples of the random variables  $Y_A$  and  $Y_B$ , yields obtained from each of two competing chemical processes. Specific values were obtained for the sample means in Chapter 1 as  $\bar{y}_A = 75.52$  and  $\bar{y}_B = 72.47$ . Consider the proposition that the  $Y_A$  data is from a normal population with the distribution  $N(75.5, 1.5^2)$ .

- (i) Determine the sampling distribution for  $\bar{Y}_A$  and from it compute  $P(75.0 \leq \bar{Y}_A \leq 76.0)$ .
- (ii) Consider the proposition that there is no difference between the yields obtainable from each process. If this is so, then the  $Y_B$  data should also be from the same normal population as that specified for  $Y_A$ . Using the sampling distribution obtained in (i), compute  $P(\bar{Y}_B \leq 72.47)$ . Comment on what this implies about the plausibility of this proposition.
- (iii) Using the sampling distribution in (i), determine the value of  $\eta_0$  for which

$$P(\bar{Y}_A \leq \eta_0) = 0.05$$

Compare this value with the computed value,  $\bar{y}_B = 72.47$ . What does this result imply about the plausibility of the proposition that both data sets come from the same population with the same distribution?

**13.29** Refer to Application Problem 13.28; this time consider an alternative proposition that in fact the  $Y_B$  is from a normal population with the distribution  $N(72.5, 2.5^2)$ . Determine the sampling distribution for  $\bar{Y}_B$  and from it compute  $P(\bar{Y}_B \leq 72.47)$  as well as  $P(72.0 \leq Y_B \leq 73.0)$ . Comment on what these results imply about the plausibility of this alternative proposition.

**13.30** A manufacturer of 10 mm diameter ball bearings uses a process which, when operating properly, is calibrated to produce ball bearings with mean diameter  $\mu = 10.00$  mm and standard deviation  $\sigma = 0.10$  mm. In order to evaluate the performance of the process at a particular point in time, a random sample of  $n$  ball bearings is to be taken and the diameters determined in a quality control laboratory. Determine

the sample size  $n$  such that

$$P(10.00 - 1.96\sigma_{\bar{X}} \leq \bar{X} \leq 10.00 + 1.96\sigma_{\bar{X}}) = 0.05$$

where  $\sigma_{\bar{X}}$  is the standard deviation of the sample mean,  $\bar{X}$ . State whatever assumptions you may have to make in answering this question.

**13.31** Refer to Application Problems 13.30. Consider that the standard practice is for the quality control lab to select a sample of  $n_s$  ball bearings, compute the specific value for the sample mean,  $\bar{x}$ , and plot it on a chart with the following characteristics: a center line representing  $\mu = 10.00$ ; an upper limit line set at  $(10 + 3\sigma/\sqrt{n_s})$  and a lower limit set at  $(10 - 3\sigma/\sqrt{n_s})$ . The process is deemed to be performing “as expected” if the value obtained for  $\bar{x}$  falls within the limits.

- (i) For a sample of 4 ball bearings, where are the upper and lower limits lines located? What is the probability of  $\bar{x}$  falling outside these limits when the process is in fact operating “as expected”.
- (ii) If a process “disturbance” shifted the true mean diameter for the manufactured ball bearings to 10.10 mm, what is the probability of detecting this shift when the result obtained from the next sample of 4 ball bearings is analyzed?

State any assumptions needed to answer these questions.

**13.32** The sample variance for the yield data presented in Chapter 1 and in Application Problem 13.28 is given  $s_A^2 = 2.05$  for process A, and  $s_B^2 = 7.62$  for process B. Consider the proposition that the  $Y_A$  data is from a normal population with the distribution  $N(75.5, 1.5^2)$ .

- (i) Determine  $P(S_A^2 \geq 1.5^2)$ ;
- (ii) If it is postulated that the  $Y_B$  data is from the same population as the  $Y_A$  data, determine the probability of overestimating the sample variance  $S_B^2$  by as much as, or worse than, the obtained value of  $s_B^2 = 7.62$ . Interpret your result in relation to the plausibility of this postulate.
- (iii) Consider the alternative postulate that the  $Y_B$  data actually came from a normal population with the distribution  $N(72.5, 2.5^2)$ . Now recompute the probability of overestimating the sample variance  $S_B^2$  by as much as, or worse than, the obtained value of  $s_B^2 = 7.62$ . Interpret this new result in relation to the plausibility of this alternative postulate.

**13.33** Refer to Application Problem 13.32 where the sample variances are given as  $s_A^2 = 2.05$  for process A, and  $s_B^2 = 7.62$  for process B. Now consider the postulate that the two sets of samples are random samples from the *same* normal population with the same, but unspecified variance. Determine the probability that a sample variance  $S_B^2$  for process B will exceed that for process A by as much as the values observed in this specific sample, or more. Comment on the implications of this result on the plausibility of this proposition.

**13.34** Random samples of size 10 each are taken from large groups of trainees instructed by Method A and Method B, and each trainee’s score on an appropriate achievement test is shown below.

|          |    |    |    |    |    |    |    |    |    |    |
|----------|----|----|----|----|----|----|----|----|----|----|
| Method A | 71 | 75 | 65 | 69 | 73 | 66 | 68 | 71 | 74 | 68 |
| Method B | 72 | 77 | 84 | 78 | 69 | 70 | 77 | 73 | 65 | 75 |

Consider the postulate that these data came from the same normal population with mean  $\mu = 70$  but whose variance is unspecified.

- (i) If this is true, what is the probability that the mean of any random sample of trainee scores will exceed 74? Interpret this result in light of individual sample means of Method A and Method B scores. How plausible is this postulate?
- (ii) Now consider an alternative postulate that scores obtained by trainees instructed by Method B are actually drawn from a normal population with mean  $\mu_B = 75$ . Determine the limits of the interval  $[(75 - 2s_B/\sqrt{10}) \leq \bar{X} \leq (75 + 2s_B/\sqrt{10})]$  and the probability that the mean score of any other random sample of 10 from this population of trainees instructed by Method B will fall into this interval. Where does the value obtained for the sample mean  $\bar{X}_A$  lie in relation to this interval? Discuss the implications of these results on the plausibility of this new postulate.

# **Chapter 14**

---

## ***Estimation***

|        |   |     |
|--------|---|-----|
| 14.1   | Introductory Concepts .....   | 488 |
| 14.1.1 | An Illustration .....   | 488 |
| 14.1.2 | Problem Definition and Key Concepts .....                           | 489 |
| 14.2   | Criteria for Selecting Estimators .....                             | 490 |
| 14.2.1 | Unbiasedness .....  | 490 |
| 14.2.2 | Efficiency .....  | 491 |
| 14.2.3 | Consistency .....   | 492 |
| 14.3   | Point Estimation Methods .....                                      | 493 |
| 14.3.1 | Method of Moments .....   | 493 |
| 14.3.2 | Maximum Likelihood .....  | 496 |
|        | Maximum Likelihood Estimate of Gaussian Population Parameters ..... | 499 |
|        | Important Characteristics of MLEs .....                             | 501 |
| 14.4   | Precision of Point Estimates .....                                  | 503 |
| 14.5   | Interval Estimates .....  | 506 |
| 14.5.1 | General Principles .....  | 506 |
| 14.5.2 | Mean of a Normal Population; $\sigma$ Known .....                   | 507 |
| 14.5.3 | Mean of a Normal Population; $\sigma$ Unknown .....                 | 508 |
| 14.5.4 | Variance of a Normal Population .....                               | 509 |
| 14.5.5 | Difference of Two Normal Populations Means .....                    | 512 |
| 14.5.6 | Interval Estimates for Parameters from other Populations .....      | 514 |
|        | Means; Large Samples .....  | 514 |
|        | Means; Small Samples .....  | 515 |
| 14.6   | Bayesian Estimation .....   | 518 |
| 14.6.1 | Background .....  | 518 |
| 14.6.2 | Basic Concept .....   | 519 |
| 14.6.3 | Bayesian Estimation Results .....                                   | 520 |
| 14.6.4 | A Simple Illustration .....   | 521 |
|        | Data .....  | 521 |
|        | The Prior Distribution .....  | 521 |
|        | The Posterior Distribution and Point Estimates .....                | 522 |
| 14.6.5 | Discussion .....  | 524 |
|        | The Bayesian Controversy: Subjectivity .....                        | 524 |
|        | Recursive Bayesian Estimation .....                                 | 525 |
|        | Choosing Prior Distributions .....                                  | 525 |
|        | Computational Issues .....  | 526 |
| 14.7   | Summary and Conclusions .....                                       | 527 |
|        | REVIEW QUESTIONS .....  | 528 |
|        | EXERCISES .....   | 530 |
|        | APPLICATION PROBLEMS .....  | 537 |

*Life is the art of drawing sufficient conclusions  
from insufficient premises*

Samuel Butler (1835–1902)

With the sampling theory foundation now firmly in place, we are finally in a position to begin building the two-tiered statistical inference edifice, starting with the first tier, Estimation, in this chapter, and finishing with Hypothesis Testing in the next chapter. The focus in the first half of this chapter is on *how to determine*, from incomplete information in sample data, unknown population parameter values needed to complete the characterization of the random variable with the pdf  $f(x|\theta)$ . The focus in the second complementary half is on *how to quantify* the unavoidable uncertainty arising from the variability in finite samples. Just as estimation theory relies on sampling theory, so does the theory of hypothesis testing rely on both estimation theory and sampling theory; the material in this chapter therefore also serves as an important link in the statistical inference chain.

---

## 14.1 Introductory Concepts

### 14.1.1 An Illustration

Consider an opinion pollster who states that 75% of undergraduate chemical engineering students in the United States prefer “closed-book” exams to “opened-book” ones, and adds a “margin of error” of  $\pm 8.5\%$  to this statement. It is instructive to begin our discussion by looking into how such information is obtained and how such statements are really meant to be interpreted.

First, in the strictest possible sense of the formal language of statistics, the “population” of concern is the *opinion* of *all* undergraduate chemical engineering students in the United States. However, in this case, many often—but somewhat imprecisely—consider the population as the students themselves (perhaps because of how this aligns with the more prevalent sociological concept of “population”). Either way, observe that we are dealing with a technically finite population (there is, after all, an actual, finite and countable number of individuals and their opinions). Practically, however, the size of this population is quite large and it is difficult (and expensive) to obtain the opinion of every single individual in this group. The pollster simply contacts a pre-determined number of subjects selected at random from this group, and asks for their individual opinions regarding the issue at stake: preference for closed-book versus opened-book exams. The premise is that there is a true, but unknown, proportion,  $\theta_c$ , that prefers closed-book exams; and that results obtained from a sample of size  $n$  can be used to deduce what  $\theta_c$  is likely to be. Next, suppose that out of 100 respondents, 75 indicated a preference for closed-book exams with the remaining 25 opting for the only other alternative. The main conclusion stated above therefore seems intuitively “reasonable” since indeed this sample shows 75 out of 100 expressing a preference for closed-book exams. But we know that sampling a different set of 100 students will quite

likely produce a different set of results. This possibility is captured by the added “margin of error”  $\pm 8.5\%$ .

This illustration raises some fundamental questions: Even though intuitive, on what basis is this “analysis” of the survey data considered as “reasonable”? How was the margin of error determined? In general, how does one determine unknown population parameters for problems that may not necessarily be as intuitive as this one? Providing answers to such questions is the objective of this chapter.

#### 14.1.2 Problem Definition and Key Concepts

*Estimation* is the process by which information about the value of a population parameter (such as  $\theta_c$  in the opinion survey above) is extracted from sample data. Because estimation theory relies heavily on sampling theory, the samples used to provide population parameter estimates are required to be random samples drawn from the population of interest. As we show shortly, this assumption significantly simplifies estimation.

There are two aspects of estimation:

1. *Point Estimation*: the process for obtaining a single “best value” for a population parameter;
2. *Interval Estimation*: the process by which one obtains a range of values that will include the true parameter, along with an appended degree of “confidence.”

Thus, in terms of the opinion poll illustration above, the point estimate of  $\theta_c$  is given as  $\hat{\theta}_c = 0.75$ . We have introduced the “hat” notation  $\hat{\cdot}$  to differentiate an estimate from the true but unknown parameter). On the other hand, the interval estimate, will be rendered as  $\hat{\theta}_c = 0.75 \pm 0.085$ , or  $0.665 < \hat{\theta}_c < 0.835$ , to which should be appended “with 95% confidence” (even though this latter appendage is usually missing in statements made for the public press).

The problem at hand may now be formulated as follows: A random variable  $X$  has a pdf  $f(x; \boldsymbol{\theta})$ , whose form is known but the parameters it contains,  $\boldsymbol{\theta}$ , are unknown; to be able to analyze  $X$  properly,  $f(x; \boldsymbol{\theta})$  needs to be completely specified, in the sense that the parameter set,  $\boldsymbol{\theta}$ , must be determined. This is done by inferring the value of the parameter vector,  $\boldsymbol{\theta}$ , from sample data, specifically, from  $\{x_1, x_2, \dots, x_n\}$ , specific values of a random sample,  $X_1, X_2, \dots, X_n$ , drawn from the population with the pdf  $f(x; \boldsymbol{\theta})$ .

The following are four key concepts that are central to estimation theory:

1. *Estimator*: Any statistic  $U(X_1, X_2, \dots, X_n)$  used for estimating the unknown quantity  $\boldsymbol{\theta}$ , or  $g(\boldsymbol{\theta})$ , a function thereof;
2. *Point Estimate*: Actual observed value  $u(x_1, x_2, \dots, x_n)$  of the estimator using specific observations  $x_1, x_2, \dots, x_n$ ;

3. *Interval Estimator:* Two statistics,  $U_L(X_1, X_2, \dots, X_n) < U_R(X_1, X_2, \dots, X_n)$ , such that  $\{U_L(X_1, X_2, \dots, X_n), U_R(X_1, X_2, \dots, X_n)\}$  represents an interval that will contain the unknown  $\theta$  (or  $g(\theta)$ ), with a probability that can be computed;
  4. *Interval Estimates:* Actual values,  $u_L(x_1, x_2, \dots, x_n)$  and  $u_R(x_1, x_2, \dots, x_n)$ , determined for the respective interval estimators from specific observations  $x_1, x_2, \dots, x_n$ .
- 

## 14.2 Criteria for Selecting Estimators

By definition, estimators are “statistics” used to estimate unknown population parameters,  $\theta$ , from actual observations. Before answering the question: how are estimators (and estimates) determined? we wish to consider first how to evaluate estimators. In particular, we will be concerned with what makes a “good” estimator, and what properties are desirable for estimators.

### 14.2.1 Unbiasedness

A statistic  $U(X_1, X_2, \dots, X_n)$  is said to be an unbiased estimator of  $g(\theta)$  if

$$E[U(X_1, X_2, \dots, X_n)] = g(\theta). \quad (14.1)$$

We know intuitively that this is a desirable property; it means, roughly, that on average, the estimator will produce an *accurate* estimate of the unknown parameter.

**Example 14.1: UNBIASED ESTIMATORS OF POPULATION MEAN**

Given a random sample  $X_1, X_2, \dots, X_n$  from a population with an unknown mean,  $\mu$ ,

(1) show that the sample average,

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad (14.2)$$

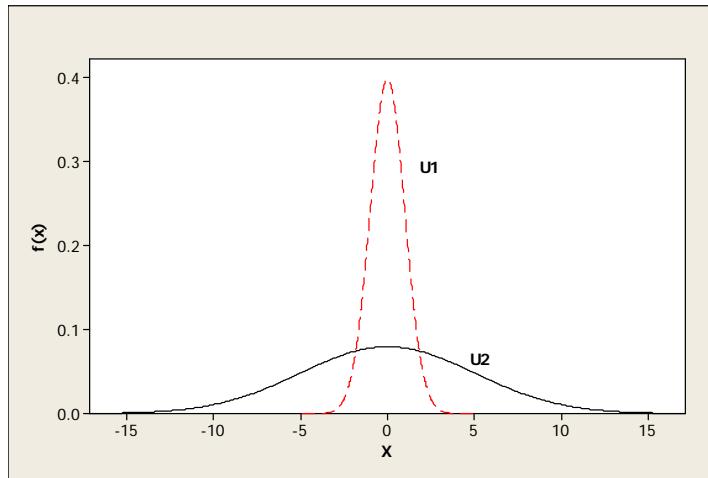
is an unbiased estimator for  $\mu$ .

(2) Also show that any other weighted average defined as

$$\bar{X}_\omega = \sum_{i=1}^n \omega_i X_i \quad (14.3)$$

is also unbiased for  $\mu$ , so long as

$$\sum_{i=1}^n \omega_i = 1 \quad (14.4)$$



**FIGURE 14.1:** Sampling distribution for the two estimators  $U_1$  and  $U_2$ :  $U_1$  is the more efficient estimator because of its smaller variance

**Solution:**

(1) By definition of the expectation, we obtain from Eq (14.2):

$$E[\bar{X}] = \frac{1}{n} \sum_{i=1}^n E[X_i] = \mu, \quad (14.5)$$

establishing that  $\bar{X}$  is indeed unbiased for  $\mu$ .

(2) By the same token, we see that

$$E[\bar{X}_\omega] = \sum_{i=1}^n \omega_i E[X_i] = \mu \sum_{i=1}^n \omega_i = \mu \quad (14.6)$$

provided that  $\sum_{i=1}^n \omega_i = 1$  as required.

### 14.2.2 Efficiency

If  $U_1(X_1, X_2, \dots, X_n)$  and  $U_2(X_1, X_2, \dots, X_n)$  are both unbiased estimators for  $g(\theta)$ , then  $U_1$  is said to be a more efficient estimator if

$$\text{Var}(U_1) < \text{Var}(U_2) \quad (14.7)$$

See Fig 14.1. The concept of efficiency roughly translates as follows: because of uncertainty, estimates produced by  $U_1$  and  $U_2$  will vary around the true value; however, estimates produced by  $U_1$ , the more efficient estimator will cluster more tightly around the true value than estimates produced by  $U_2$ . To understand why this implies greater efficiency, consider a symmetric interval of arbitrary width,  $\pm\delta$ , around the true value  $g(\theta)$ . Out of 100 estimates

produced by each estimator, because of the smaller variance associated with its sampling distribution, on average, the proportion of the estimates that will fall into this region will be higher for  $U_1$  than for  $U_2$ . The higher percentage of the  $U_1$  estimates falling into this region is a measure of the higher estimation efficiency. Thus estimates produced by  $U_1$  will, on average, be closer to the “truth” in absolute value than a corresponding number of estimates produced by  $U_2$ .

There are cases of practical importance for an unbiased estimator based on a random sample of size  $n$ , (say  $U(n)$ ) drawn from a population with pdf  $f(x; \theta)$  where there exists a smallest achievable variance. Under certain regularity conditions, we have the following result:

$$Var[U(n)] \geq \frac{\left(\frac{\partial g(\theta)}{\partial \theta}\right)^2}{nE\left\{\left[\frac{\partial \ln f(x; \theta)}{\partial \theta}\right]^2\right\}}, \quad (14.8)$$

generally known as the Cramér-Rao inequality, with the quantity on the RHS known as the Cramér-Rao (C-R) lower bound. The practical implication of this result is that no unbiased estimator  $U(n)$  can have variance lower than the C-R lower bound. An estimator with the minimum variance of all unbiased estimators (whether it achieves the C-R lower bound or not) is called a *Minimum Variance Unbiased Estimator* (MVUE).

Of the estimators in Example 14.1,  $\bar{X}$  is more efficient than  $\bar{X}_\omega$ ; in fact, it can be shown that  $\bar{X}$  is the most efficient of all unbiased estimators of  $\mu$ .

### 14.2.3 Consistency

By now, we are well aware that samples are finite subsets of populations from which they are drawn; and, as a result of unavoidable sampling variability, specific estimates obtained from various samples from the same population will not exactly equal the true values of the unknown population parameters they are supposed to estimate. Nevertheless, it would be desirable that as the sample size increases, the resulting estimates will become progressively closer to the true parameter value, until the two ultimately coincide as the sample size becomes infinite.

Mathematically, a sequence of estimators,  $U_n(\mathbf{X}), n = 1, 2, \dots$ , where  $n$  is the sample size, is said to be a consistent estimator of  $g(\theta)$  if

$$\lim_{n \rightarrow \infty} P(|U_n(\mathbf{X}) - g(\theta)| < \epsilon) = 1 \quad (14.9)$$

for every  $\epsilon > 0$ . According to this definition, a consistent sequence of estimators will produce an estimate sufficiently close to the true parameter value if the sample size is large enough.

Recall the use of Chebyshev’s inequality in Chapter 8 to establish the (weak) law of large numbers, specifically: that the relative frequency of success (the number of successes observed per  $n$  trials) approaches the actual

probability of success,  $p$ , as  $n \rightarrow \infty$ , *with probability 1*. This statement may now be interpreted as implying that the ratio  $X/n$ , where  $X$  is the binomial total number of “successes” observed in  $n$  Bernoulli trials, constitutes a consistent sequence of estimates of the population probability of success,  $p$ . This statement can be extended to sample means in general: sample means  $\sum_{i=1}^n X_i/n$  constitute a consistent sequence of estimators for the population mean when the pdf has finite variance (See Exercise 14.10).

We now proceed to discuss various ways by which to obtain point estimators.

### 14.3 Point Estimation Methods

#### 14.3.1 Method of Moments

If  $f(x; \boldsymbol{\theta})$  is the pdf of a random variable,  $X$ , with unknown parameters  $\boldsymbol{\theta}$ , then as stated in Chapter 4,  $m_i$ , the theoretical  $i^{th}$  ordinary moment of  $X$ , defined by:

$$m_i = E[X^i] \quad (14.10)$$

will be a known function of  $\boldsymbol{\theta}$ , say,  $m_i(\boldsymbol{\theta})$ . The method of moments entails obtaining from a random sample,  $X_1, X_2, \dots, X_n$ , the  $k^{th}$  sample moment

$$M_k = \frac{1}{n} \sum_{i=1}^n X_i^k \quad (14.11)$$

for  $k = 1, 2, \dots$ , as needed, equating each to the corresponding theoretical moment, and determining the values of the unknown parameters required to achieve moment matching. Solving the resulting equations,

$$M_i = m_i(\boldsymbol{\theta}) \quad (14.12)$$

for  $\boldsymbol{\theta}$  in terms of the random sample  $X_1, X_2, \dots, X_n$ , will yield:

$$\hat{\boldsymbol{\theta}} = h(X_1, X_2, \dots, X_n); \quad (14.13)$$

known as the vector of method of moment estimators. The following examples illustrate these concepts.

#### **Example 14.2: METHOD OF MOMENT ESTIMATION: EXPONENTIAL DISTRIBUTION**

Let  $X_1, X_2, \dots, X_n$  be a random sample from the exponential distribution with pdf

$$f(x; \beta) = \frac{1}{\beta} e^{-x/\beta} \quad (14.14)$$

Estimate  $\beta$ , the unknown parameter from this random sample using the

method of moments.

**Solution:**

Since there is only one parameter to be estimated, only one moment equation is required. Let us therefore choose the first moment, which, by definition, is

$$m_1 = \mu = E[X] = \beta \quad (14.15)$$

The sample analog of this theoretical moment is

$$M_1 = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (14.16)$$

and upon equating (14.15) and (14.16), we obtain:

$$\hat{\beta} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (14.17)$$

where the “hat” has been introduced to indicate an estimate and distinguish it from its true but unknown value.

Thus, the method of moments estimator for the exponential parameter  $\beta$  is:

$$U_{MM}(X_1, X_2, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i \quad (14.18)$$

When specific data sets are obtained, specific estimates of the unknown parameters are obtained from the estimators by substituting the observations  $x_1, x_2, \dots, x_n$  for the random variables  $X_1, X_2, \dots, X_n$ , as illustrated in the following examples.

**Example 14.3: SPECIFIC METHOD OF MOMENT ESTIMATES: EXPONENTIAL DISTRIBUTION**

The waiting time (in days) until the occurrence of a recordable safety incident in a certain company’s manufacturing site is known to be an exponential random variable with an unknown parameter  $\beta$ . In an attempt to improve its safety record, the company embarked on a safety performance characterization program which involves, among other things, tracking the time in between recordable safety incidents.

(1) During the first year of the program, the following data set was obtained:

$$S_1 = \{16, 1, 9, 34, 63, 44, 1, 63, 41, 29\} \quad (14.19)$$

which translates as follows: 16 days elapsed before the first recordable event occurred; 1 day thereafter the second event occurred; the third occurred 9 days after, and the fourth, 34 days after, etc. From this data record, obtain a method of moments estimate of the parameter  $\beta$ , the mean time between safety incidents.

(2) The data record for the second year of the program is:

$$S_2 = \{35, 26, 16, 23, 54, 13, 100, 1, 30, 31\} \quad (14.20)$$

Obtain the method of moment estimate of the parameter  $\beta$  for the

second-year safety performance.

**Solution:**

(1) From the results in Example 14.2, especially, Eq (14.17), the required estimate is obtained as:

$$\hat{\beta} = \bar{X} = (16 + 1 + 9 + \dots + 29)/10 = 30.1 \quad (14.21)$$

implying an average of 1 incident per  $\sim 30$  days.

(2) From the second year data set, we obtain

$$\hat{\beta} = \bar{X} = (35 + 26 + 16 + \dots + 31)/10 = 32.9 \quad (14.22)$$

At this point, it is not certain whether the difference between the two estimates is due primarily to random variability in the sample or not (the sample size,  $n = 10$  is quite small). Is it possible that a truly *significant* change has occurred in the company's safety performance and this is being reflected in the slight improvement in the average waiting time to the occurrence of recordable incidents observed in the second-year data? Hypothesis testing, the second aspect of statistical inference, provides tools for answering such questions objectively.

When there are additional parameters to estimate, the number of moments required for the estimation must naturally increase to match the number of unknown parameters, as we illustrate with the next example.

**Example 14.4: METHOD OF MOMENT ESTIMATION:  
GAUSSIAN DISTRIBUTION**

Let  $X_1, X_2, \dots, X_n$  be a random sample from a Gaussian distribution with unknown mean  $\mu$  and variance  $\sigma^2$ . Estimate these unknown parameters by the method of moments, using information from this random sample.

**Solution:**

Let  $(\theta_1, \theta_2) = (\mu, \sigma)$ . Because there are two unknown parameters, we need two moment equations to determine them. The theoretical equations for the first two moments are:

$$m_1 = E[X] = \mu \quad (14.23)$$

$$m_2 = E[X^2] = \sigma^2 + \mu^2 \quad (14.24)$$

where this last equation merely recalls the fact that  $\sigma^2 = E[X^2] - (E[X])^2$ . The sample analogs to these theoretical moment equations are:

$$M_1 = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X} \quad (14.25)$$

$$M_2 = \frac{1}{n} \sum_{i=1}^n X_i^2 \quad (14.26)$$

And now, by equating corresponding theoretical and sample moment equations and solving for the unknown parameters, we obtain, first for  $\mu$ , the estimator

$$U_1 = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X} \quad (14.27)$$

and for  $\sigma$ , the estimator

$$U_2 = \sqrt{\left( \frac{1}{n} \sum_{i=1}^n X_i^2 \right) - (\bar{X})^2} \quad (14.28)$$

Once again, given any specific observations, we obtain the actual estimates corresponding to the observations by substituting the data  $\{x_1, x_2, \dots, x_n\}$  into these estimator equations.

#### Remarks:

- Method of moments estimators are *not unique*. In Example 14.2, we could have used the second moment instead of the first, so that the theoretical moment equation would have been

$$E[X^2] = Var[X] + (E[X])^2 = 2\beta^2 \quad (14.29)$$

As such, upon equating this to the sample moment and solving, we would have obtained:

$$\hat{\beta} = \sqrt{\frac{1}{2n} \sum_{i=1}^n X_i^2} \quad (14.30)$$

which, in general, will not equal the  $\bar{X}$  prescribed in Eq (14.17).

- Thus, we cannot really talk about *the* method of moments estimator, not even *a* method of moments estimator. Note that we could just as easily have based this method on moments about the mean instead of the ordinary moments (about the origin).
- Nevertheless, these estimators are consistent (under most reasonable conditions) in the sense that empirical moments converge (in probability) to the corresponding theoretical moments.

#### 14.3.2 Maximum Likelihood

The method of maximum likelihood for estimating unknown population parameters is best illustrated with a specific example first, before generalizing.

Consider a random sample,  $X_1, X_2, \dots, X_n$ , drawn from a population possessing a Poisson distribution with unknown parameter,  $\lambda$ . By definition of a

random sample, the  $n$  random variables,  $X_i; i = 1, 2, \dots, n$  are all mutually stochastically independent; they also all have the same pdf:

$$f(x_i|\lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \quad (14.31)$$

As such, the joint distribution of the random sample is:

$$f(x_1, x_2, \dots, x_n|\lambda) = \left(\frac{e^{-\lambda}\lambda^{x_1}}{x_1!}\right) \left(\frac{e^{-\lambda}\lambda^{x_2}}{x_2!}\right) \cdots \left(\frac{e^{-\lambda}\lambda^{x_n}}{x_n!}\right) \quad (14.32)$$

which simplifies to

$$f(x_1, x_2, \dots, x_n|\lambda) = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}}{x_1!x_2!\cdots x_n!} \quad (14.33)$$

We may now note the following points about Eq (14.33):

1. If we know the parameter  $\lambda$ , then from Eq (14.33) we can obtain the probability that  $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$  jointly for various specific values of  $x_1, x_2, \dots, x_n$ . But  $\lambda$  is *unknown*, so that whatever probability is calculated from this equation will be a function of  $\lambda$ ;
2. For any particular value of  $\lambda$ , one obtains a specific value for the probability that  $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$  jointly, which corresponds to this specified value of  $\lambda$ . Thus, once the actual observations  $(x_1, x_2, \dots, x_n)$  are made, and the observed values introduced into Eq (14.33), the resulting expression becomes a function of the unknown parameter,  $\lambda$ ; it is then written as:

$$L(\lambda) = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}}{x_1!x_2!\cdots x_n!} \quad (14.34)$$

and called the *likelihood function* of the sample.

3. Eq (14.34) is called the likelihood function because it represents the likelihood of observing the specific outcome  $(x_1, x_2, \dots, x_n)$  for different values of the parameter  $\lambda$ . This is to distinguish it from the joint pdf which gives the probability of observing  $(x_1, x_2, \dots, x_n)$  when the value of  $\lambda$  is given and fixed.

Thus, with the joint pdf of a collection of random variables,  $X_1, X_2, \dots, X_n$ , (i.e.  $f(x_1, x_2, \dots, x_n|\theta)$ ), the pdf parameter vector,  $\theta$ , is given, but not the specific observations  $x_1, x_2, \dots, x_n$ ; with the likelihood function, the specific observations are given, the parameter vector is not.

Let us return to the Poisson random sample example and pose the following question: how likely is it that a population with a specific parameter  $\lambda$  produced the particular observed sample  $(x_1, x_2, \dots, x_n)$ ? Note that this is the very question that the likelihood function  $L(\lambda)$  in Eq (14.34) provides

answers to, for all conceivable values of  $\lambda$ . It now seems entirely reasonable to seek a value of  $\lambda$  (say  $\hat{\lambda}$ ) that maximizes the likelihood of observing the values  $(x_1, x_2, \dots, x_n)$  and use it as an estimate of the unknown population parameter. Such an estimate is known as the *maximum likelihood* estimate. The interpretation is that of all possible values one could entertain for the unknown parameter  $\lambda$ , this particular value,  $\hat{\lambda}$ , yields the highest possible probability of obtaining the observations  $(x_1, x_2, \dots, x_n)$ , when  $\hat{\lambda}$  is specified as the population parameter. Thus,  $\hat{\lambda}$  maximizes the likelihood of observing  $(x_1, x_2, \dots, x_n)$  in the sense that the population with the parameter  $\lambda = \hat{\lambda}$  is most likely to have produced the specific observations  $(x_1, x_2, \dots, x_n)$  (or equivalently: the specific observations at hand,  $(x_1, x_2, \dots, x_n)$ , are most likely to have come from a population for which  $\lambda = \hat{\lambda}$ ).

Determining  $\hat{\lambda}$  from Eq (14.34) requires satisfying the familiar differential calculus condition:

$$\frac{\partial L}{\partial \lambda} = 0 \quad (14.35)$$

However, the algebra involved is considerably simplified by the fact that  $L(\lambda)$  and  $\ell(\lambda) = \ln\{L(\lambda)\}$ , the so-called log-likelihood function, have the same maximum. Thus,  $\hat{\lambda}$  is often determined by maximizing the log-likelihood function instead. In this specific case, we obtain, from Eq (14.34),

$$\ell(\lambda) = \ln\{L(\lambda)\} = -n\lambda + \left( \sum_{i=1}^n x_i \right) \ln \lambda - \sum_{i=1}^n \ln x_i! \quad (14.36)$$

Differentiating with respect to  $\lambda$  and setting the result equal to zero yields:

$$\frac{\partial \ell(\lambda)}{\partial \lambda} = -n + \frac{\sum_{i=1}^n x_i}{\lambda} = 0 \quad (14.37)$$

which gives the final solution:

$$\hat{\lambda} = \frac{\sum_{i=1}^n x_i}{n}, \quad (14.38)$$

a reassuringly familiar result, since  $\lambda$  by definition, is the mean of the Poisson pdf.

The maximum likelihood *estimate* of the Poisson parameter  $\lambda$  is therefore the value  $\hat{\lambda}$  shown in Eq (14.38); the maximum likelihood *estimator* for  $\lambda$  in terms of the random sample  $X_1, X_2, \dots, X_n$  is therefore:

$$U(X_1, X_2, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i \quad (14.39)$$

The foregoing result may now be generalized as follows:

**Maximum Likelihood Estimate:** Given  $X_1, X_2, \dots, X_n$ , a random sample from a population whose pdf (continuous or discrete),  $f(x; \theta)$ , contains a vector of unknown characteristic parameters,  $\theta$ ; the likelihood function for this sample is given by:

$$L(\theta) = f(x_1; \theta)f(x_2; \theta) \cdots f(x_n; \theta) \quad (14.40)$$

the joint pdf of  $X_1, X_2, \dots, X_n$  written as a function of the unknown  $\theta$ . The value  $\hat{\theta}$  that maximizes  $L(\theta)$  is known as the maximum likelihood estimate (MLE) of  $\theta$ . (The same value  $\hat{\theta}$  maximizes  $\ell(\theta) = \ln\{L(\theta)\}$ .)

This general result is now illustrated below with several specific examples.

**Example 14.5: MAXIMUM LIKELIHOOD ESTIMATE OF AN EXPONENTIAL DISTRIBUTION PARAMETER**

Let  $X_1, X_2, \dots, X_n$  be a random sample from an exponential population with pdf

$$f(x; \beta) = \frac{1}{\beta} e^{-x/\beta} \quad (14.41)$$

Obtain the maximum likelihood estimate of the unknown population parameter,  $\beta$ , from this random sample.

**Solution:**

Since each random variable in the random sample possesses the pdf in Eq (14.41), the likelihood function in this case is:

$$\begin{aligned} L(\beta) &= \left( \frac{1}{\beta} e^{-x_1/\beta} \right) \left( \frac{1}{\beta} e^{-x_2/\beta} \right) \cdots \left( \frac{1}{\beta} e^{-x_n/\beta} \right) \\ &= \frac{1}{\beta^n} e^{-(\sum_{i=1}^n x_i)/\beta} \end{aligned} \quad (14.42)$$

From here, we easily obtain:

$$\ell(\beta) = \ln L(\beta) = -n \ln \beta - \frac{1}{\beta} \sum_{i=1}^n x_i \quad (14.43)$$

so that

$$\frac{\partial \ell(\beta)}{\partial \beta} = \frac{-n}{\beta} + \frac{1}{\beta^2} \sum_{i=1}^n x_i = 0 \quad (14.44)$$

which, for  $\beta \neq 0$ , is solved to yield the desired maximum likelihood estimate as

$$\hat{\beta} = \frac{1}{n} \sum_{i=1}^n x_i, \quad (14.45)$$

another reassuringly familiar result identical to what was obtained earlier in Example 14.2.

### Maximum Likelihood Estimate of Gaussian Population Parameters

When  $X_1, X_2, \dots, X_n$  is a random sample from a Gaussian (normal) population with pdf

$$f(x_i; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\} \quad (14.46)$$

the corresponding likelihood function is:

$$\begin{aligned} L(\mu, \sigma) &= \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\} \\ &= \left(\frac{1}{2\pi}\right)^{n/2} \frac{1}{\sigma^n} \exp\left\{-\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}\right\} \end{aligned} \quad (14.47)$$

The log-likelihood function will therefore be:

$$\ell(\mu, \sigma) = \ln L(\mu, \sigma) = -\frac{n}{2} \ln 2\pi - n \ln \sigma - \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2} \quad (14.48)$$

To determine the maximum likelihood estimates of the unknown parameters  $\mu, \sigma$  requires taking appropriate derivatives in Eq (14.48), equating to zero and solving for the unknowns; i.e.,

$$\begin{aligned} \frac{\partial \ell}{\partial \mu} &= \frac{-2 \sum_{i=1}^n (x_i - \mu)}{2\sigma^2} = 0 \\ \Rightarrow \sum_{i=1}^n x_i - n\mu &= 0 \end{aligned} \quad (14.49)$$

and

$$\frac{\partial \ell}{\partial \sigma} = -\frac{n}{\sigma} + \frac{\sum_{i=1}^n (x_i - \mu)^2}{\sigma^3} = 0 \quad (14.50)$$

These two equations must now be solved simultaneously. First, from Eq (14.49), we obtain:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i \quad (14.51)$$

which, when introduced into Eq (14.50) and simplified, gives the second and final result:

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^n (x_i - \hat{\mu})^2}{n}} \quad (14.52)$$

Observe that the MLE of  $\mu$  in Eq (14.51) is the same as the sample mean, *but* the MLE of  $\sigma$  in Eq (14.52) is *not* the same as the sample standard deviation. For large  $n$ , of course, the difference becomes negligible, but this illustrates an important point: because the sample standard deviation

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} \quad (14.53)$$

(where  $\bar{x}$  is the sample mean) is unbiased for  $\sigma$ , this implies that the MLE for  $\sigma$  is biased.

### Important Characteristics of MLEs

Of the many characteristics of maximum likelihood estimates, the following are the two most important we wish to highlight:

1. *Invariance Property*: If  $\hat{\theta}$  is the MLE of the population parameter vector  $\theta$ , then  $g(\hat{\theta})$  is also the MLE of  $g(\theta)$ .
2. *Asymptotic Properties*: As  $n \rightarrow \infty$ , the MLE approaches minimum variance and is unbiased. Thus, the MLE is asymptotically unbiased, asymptotically efficient, and consistent.

Thus, according to the first property, from the MLE of the sample standard deviation in Eq (14.52), we immediately know that the MLE of the variance will be given by

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (x_i - \hat{\mu})^2}{n} \quad (14.54)$$

The second property makes large sample MLEs very attractive.

The following are a few more examples of MLEs.

#### **Example 14.6: MLE OF A BINOMIAL/BERNOULLI PROBABILITY OF “SUCCESS”**

Let  $X_1, X_2, \dots, X_n$  be the result obtained from  $n$  Bernoulli trials where the probability of “success” is  $p$ , i.e.

$$X_i = \begin{cases} 1; & \text{with probability } p \\ 0; & \text{with probability } q = 1 - p \end{cases} \quad (14.55)$$

If the random variable,  $X$ , is the total number of successes in  $n$  trials,  $\sum_i X_i$ , obtain a MLE for  $p$ .

#### **Solution:**

There are several ways of approaching this problem. The first approach is direct, from the perspective of the Bernoulli random variable; the second makes use of the fact that the pdf of a sum of  $n$  Bernoulli random variables is a Binomial random variable.

To approach this directly, we recall from Chapter 8, the compact pdf for the Bernoulli random variable

$$f(x) = p^{I_S}(1-p)^{I_F} \quad (14.56)$$

where the “success indicator”,  $I_S$ , is defined as:

$$I_S = \begin{cases} 1; & \text{for } X = 1 \\ 0; & \text{for } X = 0 \end{cases} \quad (14.57)$$

and its complement, the “failure indicator”,  $I_F$

$$I_F = \begin{cases} 1; & \text{for } X = 0 \\ 0; & \text{for } X = 1 \end{cases} \quad (14.58)$$

The joint pdf for the random sample  $X_1, X_2, \dots, X_n$  is therefore given by

$$f(x_1, x_2, \dots, x_n) = p^{\sum_{i=1}^n I_{S_i}} (1-p)^{\sum_{i=1}^n I_{F_i}} \quad (14.59)$$

and now, because  $I_{S_i} = X_i$  and  $I_F = 1 - X_i$ , then Eq (14.59) reduces to

$$f(x_1, x_2, \dots, x_n) = p^{\sum_{i=1}^n X_i} (1-p)^{n - \sum_{i=1}^n X_i} \quad (14.60)$$

so that the likelihood function is:

$$L(p) = p^{\sum_{i=1}^n X_i} (1-p)^{n - \sum_{i=1}^n X_i} \quad (14.61)$$

with the log-likelihood function,

$$\ell(p) = \sum_{i=1}^n X_i \ln p + \left( n - \sum_{i=1}^n X_i \right) \ln(1-p) \quad (14.62)$$

From here, the usual differential calculus exercise results in:

$$\begin{aligned} \frac{\partial \ell}{\partial p} &= \frac{\sum_{i=1}^n X_i}{p} - \frac{(n - \sum_{i=1}^n X_i)}{(1-p)} = 0 \\ \Rightarrow \frac{\sum_{i=1}^n X_i}{p} &= \frac{(n - \sum_{i=1}^n X_i)}{(1-p)} \end{aligned} \quad (14.63)$$

which, when solved for  $p$ , yields the result:

$$\hat{p} = \frac{\sum_{i=1}^n X_i}{n} \quad (14.64)$$

as one would intuitively expect.

The second approach uses not the joint pdf of  $X_1, X_2, \dots, X_n$ , but the related pdf of a function of  $X_1, X_2, \dots, X_n$ , i.e.  $X = X_1 + X_2 + \dots + X_n$ , which is known to be a Binomial random variable: i.e.

$$f(x) = \binom{n}{X} p^X (1-p)^{n-X} \quad (14.65)$$

so that the likelihood in this case (for  $X = \sum_{i=1}^n X_i$ ) is:

$$L(p) = \binom{n}{X} p^X (1-p)^{n-X} \quad (14.66)$$

It is a relatively simple exercise left to the reader (See Exercise 14.19) to establish that this function is maximized when

$$\hat{p} = \frac{X}{n} \quad (14.67)$$

exactly the same result as in Eq (14.64) since  $X = \sum_{i=1}^n X_i$ .

Thus, the MLE for  $p$  is the total number of successes in  $n$  trials divided by the total number of trials, as one would expect.

The next example is a specific application of this general result.

**Example 14.7: MLE OF OPINION SURVEY PARAMETER**

In the opinion survey illustration used to open this chapter, out of 100 students surveyed by the opinion pollster, 75 indicated a preference for closed-book exams. If one considers the sampling of each student to constitute a Bernoulli trial in which the outcome, “preference for closed-book exams,” is nominally considered a “success,” and if the student population is such that the true fraction with a preference for closed-book exams is  $\theta_c$ , find the MLE of  $\theta_c$  from the survey result.

**Solution:**

This is clearly a case in which  $X$ , the total number of respondents showing a preference for closed-book exams, is a Binomial random variable, in this case with  $n = 100$  and  $p = \theta_c$  unknown. With the observation,  $x = 75$ , we therefore easily obtain, from the results of Example 14.6, that

$$\hat{p} = \hat{\theta}_c = 0.75 \quad (14.68)$$

## 14.4 Precision of Point Estimates

We have shown that, given a random sample  $X_1, X_2, \dots, X_n$ , regardless of the underlying population distribution from which the sample came, the estimators

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (14.69)$$

and

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} \quad (14.70)$$

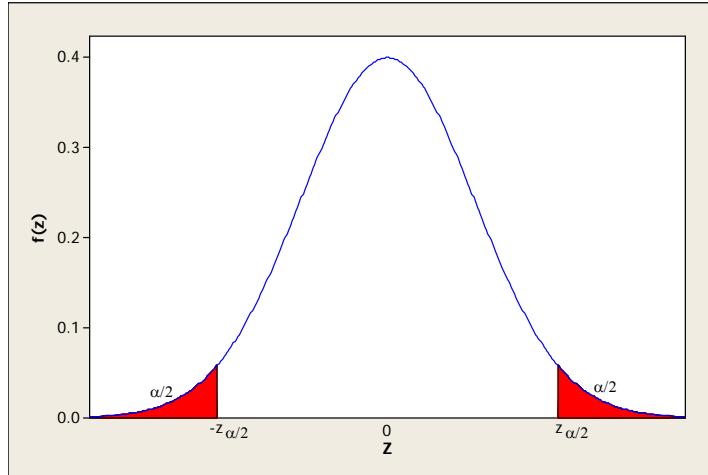
are both unbiased for the population mean  $\mu$ , and variance  $\sigma^2$ , respectively. They are thus both accurate. For any specific set of observations,  $(x_1, x_2, \dots, x_n)$ , the computed

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (14.71)$$

in general will not be identically equal to  $\mu$ ; neither will

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \quad (14.72)$$

be identical to  $\sigma^2$ , in general. The question of concern now is this: *how close to  $\mu$  do we expect  $\bar{x}$ , and  $s^2$  to  $\sigma^2$ ?* To answer this question, we need sampling



**FIGURE 14.2:** Two-sided tail area probabilities of  $\alpha/2$  for the standard normal sampling distribution

distributions that will enable us make appropriate probabilistic statements about the proximity of estimates to unknown parameter values.

We therefore begin by recalling that for large  $n$ ,  $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$  has a distribution that is approximately  $N(0, 1)$ , with the implication that:

$$P \left[ -z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2} \right] = (1 - \alpha) \quad (14.73)$$

where  $z_{\alpha/2}$  is that value of  $z$  with a tail area probability of  $\alpha/2$ , as illustrated in Fig 14.2. We may therefore state that with probability  $(1 - \alpha)$ , the proximity of  $\bar{X}$  to  $\mu$  is characterized by:

$$\frac{|\bar{X} - \mu|}{\sigma/\sqrt{n}} < z_{\alpha/2} \quad (14.74)$$

For the specific case where  $\alpha$  is chosen as 0.05, then,  $z_{\alpha/2}$ , the value of the standard normal random variable,  $z$ , for which the tail area probability is 0.025 is 1.96. As a result,

$$\frac{|\bar{X} - \mu|}{\sigma/\sqrt{n}} < 1.96 \quad (14.75)$$

implying that

$$|\bar{X} - \mu| < 1.96\sigma/\sqrt{n} \quad (14.76)$$

The implication of this result is that given  $\sigma$ , we can state that  $\bar{X}$  will not be farther away from the true value,  $\mu$ , by more than  $1.96\sigma/\sqrt{n}$ , with probability 0.95. Alternatively, Eq (14.76) may be restated as:

$$\mu = \bar{X} \pm \frac{1.96\sigma}{\sqrt{n}} \quad (14.77)$$

indicating that, with 95% confidence, the true value,  $\mu$ , will lie in the interval that stretches  $1.96\sigma/\sqrt{n}$  to the right and to the left of the sample average. We shall return to this statement shortly, but for now we note that the estimate becomes more precise with larger sample size  $n$ ; and in the limit as  $n \rightarrow \infty$ , the estimate coincides precisely with the true value.

When  $\sigma$  is unknown, and the sample size  $n$  is small, we simply replace  $\sigma$  with the estimate  $S$ , and replace  $z$  with the  $t$ -distribution equivalent in Eq (14.74) to obtain:

$$\frac{|\bar{X} - \mu|}{S/\sqrt{n}} < t_{\alpha/2}(n-1) \quad (14.78)$$

with probability  $(1 - \alpha)$ , where  $n - 1$  represents the degrees of freedom associated with the  $t$ -distribution.

For the variance, the question “how close is  $S^2$  to  $\sigma^2$  (or equivalently, how close is  $S$  to  $\sigma$ )” is answered by appealing to the result that when sampling from a normal population,  $(n-1)S^2/\sigma^2$  has a  $\chi^2(n-1)$  distribution. This theoretical sampling distribution may then be used to make probability statements about the closeness of  $S^2$  to  $\sigma^2$ . Unfortunately, not much can be said in general when the sampling is from non-normal populations.

For binomial proportions, the question “how close is  $\hat{p} = X/n$  to  $p$ ” is answered in the same manner as discussed above for the mean and the variance, provided the sample size  $n$  is large. Since the variance of the binomial random variable,  $\sigma_X^2 = np(1-p)$ , so that  $\sigma_p^2 = p(1-p)/n$ , we may use the Central Limit Theorem to infer that  $[(\hat{p}-p)/\sigma_p] \sim N(0, 1)$  and hence use the standard normal approximation to the sampling distribution of  $\hat{p}$  to make probability statements about the proximity of  $\hat{p} = X/n$  to  $p$ . Thus, for example,

$$\frac{|\hat{p} - p|}{\sigma_p/\sqrt{n}} < z_{\alpha/2} \quad (14.79)$$

with probability  $(1 - \alpha)$ , as the next example illustrates.

#### **Example 14.8: PRECISION OF OPINION SURVEY RESULT**

In Example 14.7, the MLE of  $p$ , the true proportion of college students with a preference for closed-book exams, was estimated as 0.75 from the opinion survey result of 100 students. How precise is this estimate?

##### **Solution:**

As was the case in Example 14.7, on the assumption that  $X$ , the total number of respondents with a preference for closed-book exams, is a Binomial random variable, then  $\sigma_X^2 = np(1-p)$ . And for  $\hat{p} = X/n$ ,

$$Var(\hat{p}) = \frac{\sigma_X^2}{n^2} = \frac{p(1-p)}{n} \quad (14.80)$$

so that  $\sigma_p$ , the standard deviation for  $p$ , is:

$$\sigma_p = \frac{\sqrt{p(1-p)}}{\sqrt{n}} \quad (14.81)$$

with  $p$  estimated as 0.75. Assuming that  $n = 100$  is sufficiently large so that the standard normal approximation for  $\{(\hat{p} - p)/\sigma_p\}$  holds in this case, we obtain immediately from Eq (14.77) that, with probability 0.95,

$$\theta_c = p = 0.75 \pm 1.96\sqrt{0.75 \times 0.25}/10 = 0.75 \pm 0.085. \quad (14.82)$$

This is how the survey's "margin of error" quoted at the beginning of the chapter was determined.

We may now observe that in adding a measure of precision to point estimates, the net result has appeared in the form of an interval within which the true parameter is expected to lie, with a certain pre-specified probability. This motivates the concept of interval estimation.

## 14.5 Interval Estimates

### 14.5.1 General Principles

Primarily because they are based on incomplete population information from samples (samples which are themselves subject to variability), point estimates, we now know, will never coincide identically with the theoretical parameters being estimated. In the previous section, we dealt with this problem by seeking to quantify the precision of the point estimate. We did this by determining, in a probabilistic sense, how close  $\hat{\theta}$ , the estimate, is to  $\theta$ , the true but unknown value. The results we obtained for  $\bar{X}$  as an estimator of  $\mu$  began as a probability statement in Eq (21.74), whose argument, when rearranged as follows:

$$\bar{X} - z_{\alpha/2} \left( \frac{\sigma}{\sqrt{n}} \right) < \mu < \bar{X} + z_{\alpha/2} \left( \frac{\sigma}{\sqrt{n}} \right) \quad (14.83)$$

gives the interval within which we expect the true value  $\mu$  to lie, with probability  $(1 - \alpha)$ . This provides a different way of estimating  $\mu$  — an approach that combines, in one self-contained statement, the estimate and a probabilistic measure of its precision; it is called an *interval estimate*.

There are, therefore, two main aspects of an interval estimate:

1. The boundaries of the interval; and
2. The associated probability (usually termed the "degree of confidence") that the specified random interval will contain the unknown parameter.

The interval estimators are the two statistics  $U_L$  and  $U_R$  used to determine the left and right boundaries respectively; the sampling distribution of the point

estimator is used to obtain the appropriate interval estimates that correspond to the pre-specified probability,  $(1 - \alpha)$ , the desired degree of confidence. The result is then typically known as the  $(1 - \alpha) \times 100\%$  confidence interval. Since, as discussed in Chapter 13, the nature of sampling distributions depends on what is known about the underlying population, the same is true for methods for obtaining interval estimates, and for the very same reasons.

#### 14.5.2 Mean of a Normal Population; $\sigma$ Known

If  $X_1, X_2, \dots, X_n$  is a random sample from a normal population, then we know that  $\bar{X}$ , the sample average, is a good point estimator that enjoys many desirable properties. We also know that

$$Z = \left( \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right) \quad (14.84)$$

has a standard normal,  $N(0, 1)$ , distribution. From this sampling distribution for the statistic  $\bar{X}$ , we now obtain the following probability statement:

$$P \left( -z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2} \right) = (1 - \alpha) \quad (14.85)$$

as we did earlier in Eq (21.74), which converts to the interval shown in Eq (14.83), implying finally that the interval  $[\bar{X} \pm z_{\alpha/2}(\sigma/\sqrt{n})]$  contains  $\mu$  with probability  $(1 - \alpha)$ . Specifically for  $\alpha = 0.05$ , the commonly used default value,  $z_{\alpha/2} = 1.96$ , so that the resulting interval,

$$CI_\mu = \bar{X} \pm 1.96(\sigma/\sqrt{n}), \quad (14.86)$$

is therefore the 95% confidence interval for  $\mu$ , the mean of a normal population estimated from a random sample of size  $n$ .

The general procedure for obtaining interval estimates for the mean of a normal population is therefore as follows:

1. Determine the point estimator (the sample average) and its distribution (a standard normal for the normalized average, which will include  $\sigma$ , the population variance, assumed known);
2. Determine the end points of an interval that will contain the unknown mean  $\mu$ , with specified probability (typically  $(1 - \alpha)$ , with  $\alpha = 0.05$ ).

The following example illustrates this procedure.

**Example 14.9: INTERVAL ESTIMATES FOR MEANS OF**

### PROCESS YIELDS

Given that the result of a series of 50 experiments performed on the chemical processes discussed in Chapter 1 constitute random samples from the respective populations for the yields,  $Y_A$  and  $Y_B$ , assume that these are two normal populations and obtain 95% confidence interval estimates for the population means  $\mu_A$  and  $\mu_B$ , given the respective population standard deviations as  $\sigma_A = 1.5$  and  $\sigma_B = 2.5$ .

#### Solution:

From the supplied data, we obtain the sample averages as:

$$\bar{y}_A = 75.52; \bar{y}_B = 72.47 \quad (14.87)$$

and given the standard deviations and  $n = 50$ , we obtain the following interval estimates:

$$\mu_A = 75.52 \pm 1.96(1.5/\sqrt{50}) = 75.52 \pm 0.42 \quad (14.88)$$

and

$$\mu_B = 72.47 \pm 1.96(2.5/\sqrt{50}) = 72.47 \pm 0.69 \quad (14.89)$$

The implication is that, with 95% confidence, the mean yield for each process is characterized as follows: for process A,  $75.10 < \mu_A < 75.94$ ; and for process B,  $71.78 < \mu_B < 73.16$ . As a preview of upcoming discussions, note that these two 95% confidence intervals do not overlap.

#### 14.5.3 Mean of a Normal Population; $\sigma$ Unknown

When the population standard deviation is unknown, of course the point estimate remains unchanged as  $\bar{X}$ , but now we must introduce the standard deviation estimator,

$$S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}} \quad (14.90)$$

for the unknown standard deviation. And from Chapter 13, we know that the sampling distribution of the statistic:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \quad (14.91)$$

is the  $t$ -distribution with  $\nu = n - 1$  degrees of freedom. As such, we know that:

$$P \left[ -t_{\alpha/2}(n-1) < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2}(n-1) \right] = (1 - \alpha), \quad (14.92)$$

which is the  $t$ -distribution analog to Eq (21.74). As usual,  $t_{\alpha/2}(n-1)$  is the value of the  $t$  random variable that yields a tail area probability of  $\alpha/2$  for a  $t$ -distribution with  $n-1$  degrees of freedom. Thus, when the standard deviation,

$\sigma$ , is unknown, the required  $(1 - \alpha) \times 100\%$  confidence interval for a normal population mean,  $\mu$ , is:

$$\bar{X} - t_{\alpha/2}(n-1) \left( \frac{S}{\sqrt{n}} \right) < \mu < \bar{X} + t_{\alpha/2}(n-1) \left( \frac{S}{\sqrt{n}} \right) \quad (14.93)$$

**Example 14.10: INTERVAL ESTIMATES FOR MEANS OF PROCESS YIELDS: UNKNOWN POPULATION VARIANCES**

Repeat the problem in Example 14.9 and obtain 95% confidence interval estimates for the population means  $\mu_A$  and  $\mu_B$ , still assuming that the data came from normal populations, but with *unknown* standard deviations.

**Solution:**

As obtained in the earlier example, the sample averages remain:

$$\bar{y}_A = 75.52; \bar{y}_B = 72.47 \quad (14.94)$$

From the data, we also obtain the sample standard deviations as

$$s_A = 1.43; s_B = 2.76 \quad (14.95)$$

With  $n = 50$ , so that  $\nu = 49$ , we obtain from MINITAB

$$t_{0.025}(49) = 2.01 \quad (14.96)$$

(This is obtained using the “inverse cumulative probability” feature: (**Calc > Prob Distr. > t > Inverse Cum Prob**) entering 49 for the degrees of freedom, and 0.025 as the desired tail area. This returns the result that  $P(T < -2.01) = 0.025$ , which, by symmetry implies that  $P(T > 2.01) = 0.025$ , so that  $t_{0.025}(49) = 2.01$ .)

We now easily obtain the following interval estimates:

$$\mu_A = 75.52 \pm 2.01(1.43/\sqrt{50}) = 75.52 \pm 0.41 \quad (14.97)$$

and

$$\mu_B = 72.47 \pm 2.01(2.76/\sqrt{50}) = 72.47 \pm 0.78 \quad (14.98)$$

The 95% confidence intervals for the mean yield for each process is now as follows: for Process A,  $(75.11 < \mu_A < 75.93)$ ; and for Process B,  $(71.69 < \mu_B < 73.25)$ .

Note that these interval estimates are really not that different from those obtained in Example 14.9; in fact, the estimates for  $\mu_A$  are virtually identical. There are two reasons for this: first, and foremost, the sample estimates of the population standard deviation,  $s_A = 1.43; s_B = 2.76$ , are fairly close to the corresponding population values  $\sigma_A = 1.5$  and  $\sigma_B = 2.5$ ; second, the sample size  $n = 50$  is sufficiently large so that the difference between the  $t$ -distribution and the standard normal is almost negligible (observe that  $z_{0.025} = 1.96$  is only 2.5% less than  $t_{0.025}(49) = 2.01$ ).

Again, note that these two 95% confidence intervals also do not overlap.

#### 14.5.4 Variance of a Normal Population

Obtaining interval estimates for the variance of a normal population follows the same principles outlined above: obtain the sampling distribution of an appropriate statistic (the estimator) and use it to make probabilistic statements about an interval that is expected to contain the unknown parameter. In the case of the population variance, the estimator is:

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} \quad (14.99)$$

and the sampling distribution is obtained from the fact that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1) \quad (14.100)$$

when sampling from a normal population.

From here, we are now able to make the following probability statement:

$$P \left[ \chi_{1-\alpha/2}^2(n-1) < \frac{(n-1)S^2}{\sigma^2} < \chi_{\alpha/2}^2(n-1) \right] = (1-\alpha), \quad (14.101)$$

where, because of the asymmetry of the  $\chi^2$  distribution, the left boundary is the number  $\chi_{1-\alpha/2}^2(n-1)$ , the value of the  $\chi^2$  random variable for which the area to its right under the  $\chi^2$ -distribution with  $n-1$  degrees of freedom is  $(1-\alpha/2)$ , so that the tail area to the left will be the desired  $\alpha/2$ ; on the right boundary, the value is  $\chi_{\alpha/2}^2(n-1)$ . Fig 14.3 shows such a sampling distribution with 9 degrees of freedom, along with the left and right boundary values for  $\alpha = 0.05$ . As a result of the asymmetry,  $\chi_{1-\alpha/2}^2(n-1) = 2.7$  while  $\chi_{\alpha/2}^2(n-1) = 19.0$  in this case.

The expression in Eq (14.101), when rearranged carefully, yields the result for the  $(1-\alpha) \times 100\%$  confidence interval for the population variance,  $\sigma^2$ , as:

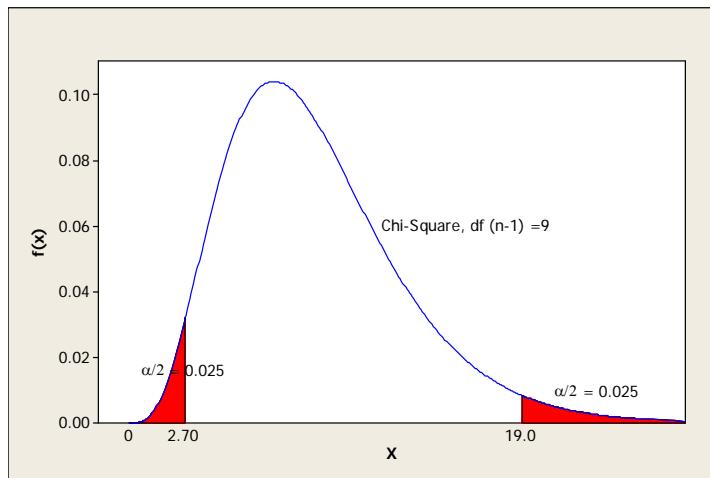
$$\frac{(n-1)S^2}{\chi_{\alpha/2}^2(n-1)} < \sigma^2 < \frac{(n-1)S^2}{\chi_{1-\alpha/2}^2(n-1)} \quad (14.102)$$

The 95% confidence interval on the standard deviation is obtained by taking square roots to yield:

$$S \sqrt{\frac{(n-1)}{\chi_{\alpha/2}^2(n-1)}} < \sigma < S \sqrt{\frac{(n-1)}{\chi_{1-\alpha/2}^2(n-1)}} \quad (14.103)$$

#### Example 14.11: INTERVAL ESTIMATES FOR VARIANCES OF PROCESS YIELDS

Obtain the 95% confidence interval estimates for the population variances  $\sigma_A^2$  and  $\sigma_B^2$  for the process yield data sets discussed in Chapter 1 and in Example 14.9. Assume, as before, that the data came from normal populations. How do the respective population variances and



**FIGURE 14.3:** Two-sided tail area probabilities of  $\alpha/2 = 0.025$  for a Chi-squared distribution with 9 degrees of freedom

standard deviations specified in Example 14.9 fit into these estimated intervals?

**Solution:**

First, we recall the sample standard deviations computed in Example 14.10 as  $s_A = 1.43$ ;  $s_B = 2.76$ ; with  $\alpha$  specified as 0.05, we obtain from MINITAB (again using the inverse cumulative probability feature, this time for the  $\chi^2$  distribution, with  $\nu = 49$ ), that:

$$\chi_{0.975}^2(49) = 31.6; \chi_{0.025}^2(49) = 70.2 \quad (14.104)$$

From here, using Eq (14.102), the following interval estimates for the variances are obtained as:

$$1.43 < \sigma_A^2 < 3.18 \quad (14.105)$$

$$5.33 < \sigma_B^2 < 11.85 \quad (14.106)$$

or, upon taking square roots, the interval estimates for the standard deviations are:

$$1.20 < \sigma_A < 1.78 \quad (14.107)$$

$$2.31 < \sigma_B < 3.44 \quad (14.108)$$

Note that the population standard deviations, specified earlier in Example 14.9 as  $\sigma_A = 1.5$  and  $\sigma_B = 2.5$ , fall entirely within their respective estimated intervals (the variances  $\sigma_A^2 = 2.25$ ;  $\sigma_B^2 = 6.25$  also fall within the respective estimated intervals for variances).

The results discussed above for interval estimates of single means and variances from normal populations have implications for hypothesis testing, as we

show in the next chapter. They can be extended to interval estimates of the differences between the means of two normal populations, as we will now do: this also has implications for hypothesis testing.

#### 14.5.5 Difference of Two Normal Populations Means

Consider  $X_1, X_2, \dots, X_n$ , a random sample from a normal population, with the distribution  $N(\mu_X, \sigma_X^2)$  and independent from  $Y_1, Y_2, \dots, Y_m$ , another random sample from a different normal population, with the distribution  $N(\mu_Y, \sigma_Y^2)$ , where the sample sizes,  $n$  and  $m$  need not be equal (i.e.  $n \neq m$ ). We are now concerned with determining

$$\delta_{XY} = \mu_X - \mu_Y \quad (14.109)$$

the difference between the two population means, by obtaining a point estimate along with a confidence interval.

If  $\bar{X}$  is the MLE for  $\mu_X$ , and  $\bar{Y}$  is the MLE for  $\mu_Y$ , then it is straightforward to show (see Exercise 14.27) that  $\bar{D}$ , defined as:

$$\bar{D} = \bar{X} - \bar{Y} \quad (14.110)$$

is the MLE for  $\delta_{XY}$ ; it is also unbiased. And now, to obtain the interval estimate for  $\bar{D}$ , we need its sampling distribution, which is obtained as follows: we know from previous results that  $\bar{X} \sim N(\mu_X, \sigma_X^2/n)$  and  $\bar{Y} \sim N(\mu_Y, \sigma_Y^2/m)$ ; and from results about distributions of sums of Gaussian random variables, we now obtain that  $\bar{D} \sim N(\delta_{XY}, v^2)$  where:

$$\delta_{XY} = \mu_X - \mu_Y \quad (14.111)$$

$$v^2 = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m} \quad (14.112)$$

The latter equation arises from the fact that

$$Var(\bar{D}) = Var(\bar{X}) + Var(\bar{Y}) \quad (14.113)$$

by independence. And now, if  $\sigma_X^2$  and  $\sigma_Y^2$  are known (so that  $v^2$  is also known) then observe that the statistic  $(\bar{D} - \delta_{XY})/v$  has a standard normal distribution. Thus

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(0, 1) \quad (14.114)$$

from which we obtain the probability statement:

$$P \left[ -z_{\alpha/2} < \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} < z_{\alpha/2} \right] = 1 - \alpha \quad (14.115)$$

so that the  $(1 - \alpha) \times 100\%$  confidence interval for  $\delta_{XY}$  is given as:

$$\delta_{XY} = (\bar{X} - \bar{Y}) \pm z_{\alpha/2} \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \quad (14.116)$$

The next example illustrates this result.

**Example 14.12: INTERVAL ESTIMATES FOR DIFFERENCE BETWEEN TWO PROCESS YIELD MEANS**

Obtain a 95% confidence interval estimate for the *difference* between the population means  $\mu_A$  and  $\mu_B$  for the process yield data in Example 14.9, given the respective population standard deviations as  $\sigma_A = 1.5$  and  $\sigma_B = 2.5$ .

**Solution:**

Since  $\bar{y}_A = 75.52$ ;  $\bar{y}_B = 72.47$ , so that  $d = \bar{y}_A - \bar{y}_B = 3.05$ , the desired 95% confidence interval is obtained from Eq (14.116) as

$$\delta_{AB} = 3.05 \pm 1.96 \sqrt{(2.25 + 6.25)/50} = 3.05 \pm 0.81 \quad (14.117)$$

Thus, "with 95% confidence," we expect  $2.24 < \delta_{AB} < 3.86$ .

The result of this example foreshadows part of the upcoming discussion in the next chapter on hypothesis testing. For now we simply note the most obvious implication: it is highly unlikely that the mean of the yield obtainable from process A is the same as that from process B; in fact, the evidence seems to support the postulate that the mean yield obtainable from process A is greater than that from process B, by as little as 2.24 and possibly by as large as 3.86.

This example also sheds light in general on how we can use the interval estimate of the difference between two means to assess the equality of two normal population means:

1. If the interval estimate for  $\mu_X - \mu_Y$  contains the number 0, the implication is that  $\mu_X$  and  $\mu_Y$  are very likely equal;
2. If the interval estimate for  $\mu_X - \mu_Y$  lies entirely to the right of 0, the implication is that, very likely,  $\mu_X > \mu_Y$ ; and finally,
3. If the interval estimate for  $\mu_X - \mu_Y$  lies entirely to the left of 0, the implication is that, very likely,  $\mu_X < \mu_Y$ .

When the population variances,  $\sigma_X^2$  and  $\sigma_Y^2$ , are unknown, in general, things become quite complicated, especially when  $n \neq m$ . Under these circumstances, it is customary to use

$$\delta_{XY} = (\bar{X} - \bar{Y}) \pm z_{\alpha/2} \sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}} \quad (14.118)$$

as an *approximate*  $(1 - \alpha) \times 100\%$  confidence interval for  $\delta_{XY}$ , where the variances in Eq (14.116) have been replaced by the sample equivalents.

When  $\sigma_X^2$  and  $\sigma_Y^2$  are unknown but equal to  $\sigma^2$ , we can use the  $t$ -distribution as we have done previously to obtain

$$\delta_{XY} = (\bar{X} - \bar{Y}) \pm t_{\alpha/2}(\nu) S_p \sqrt{\frac{1}{n} + \frac{1}{m}} \quad (14.119)$$

as the  $(1 - \alpha) \times 100\%$  confidence interval for  $\delta_{XY}$ , where  $\nu$ , the degrees of freedom is defined as:

$$\nu = n + m - 2 \quad (14.120)$$

and  $S_p$ , known as the “pooled standard deviation,” is obtained from the expression

$$\frac{(n + m - 2)S_p^2}{\sigma^2} = \frac{(n - 1)S_X^2 + (m - 1)S_Y^2}{\sigma^2} \quad (14.121)$$

or,

$$S_p = \sqrt{\frac{(n - 1)S_X^2 + (m - 1)S_Y^2}{(n + m - 2)}} \quad (14.122)$$

the positive square root of a weighted average of the two sample variances.

#### 14.5.6 Interval Estimates for Parameters from other Populations

While the most readily available results for interval estimates are for samples from Gaussian populations, it is still possible to obtain interval estimates for parameters from non-Gaussian populations. One simply needs to remember that the key to interval estimation is the sampling distribution of the estimator. If we are able to obtain the appropriate sampling distribution, it can be used to make the sort of probabilistic statements on which interval estimates are based.

##### Means; Large Samples

Fortunately, when sample sizes are large, it is possible to invoke the central limit theorem to determine that, regardless of the underlying distribution (Gaussian or not),  $(\bar{X} - \mu)/(\sigma/\sqrt{n})$  possesses an approximate  $N(0, 1)$  distribution, with the approximation improving as  $n \rightarrow \infty$ . Furthermore, even if  $\sigma$  is unknown (as is usually the case in most problems of practical relevance), the large sample size makes it acceptable to approximate  $\sigma$  with  $S$ , the sample standard deviation. Thus, no new results are required under these circumstances. The following example illustrates this point.

##### Example 14.13: INTERVAL ESTIMATE FOR MEAN OF INCLUSIONS DATA

The number of *inclusions* found on glass sheets produced in the manufacturing process discussed in Chapter 1 has been identified as a Poisson random variable with parameter  $\lambda$ . If the data in Table 1.2 is considered

a random sample of 60 observations, obtain a 95% confidence interval for the parameter  $\lambda$ .

**Solution:**

The data is from a Poisson population, not from a Gaussian one; but the sample size is large. Having determined, earlier in this chapter, that the sample mean is the MLE for the Poisson parameter, we conclude from the supplied data that  $\hat{\lambda} = \bar{x} = 1.02$ ; also the sample standard deviation,  $s = 1.1$ . A sample size of 60 is typically considered large enough ( $n > 50$ ) so that the standard normal approximation for the distribution of  $(\bar{X} - \lambda)/(\sigma/\sqrt{n})$  is valid in this case. The immediate implication is that the desired 95% confidence interval is:

$$\lambda = 1.02 \pm 1.96(1.1/\sqrt{60}) = 1.02 \pm 0.28 \quad (14.123)$$

so that, with 95% confidence, we can expect the true mean number of inclusions found on the glass sheets made in this manufacturing site to be characterized as:  $0.74 < \lambda < 1.30$

Thus, the  $(1 - \alpha) \times 100\%$  interval estimate for the binomial proportion  $p$  is obtained as

$$\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} \quad (14.124)$$

where the sample estimates of the variance have been introduced for the unknown population variance.

### Means; Small Samples

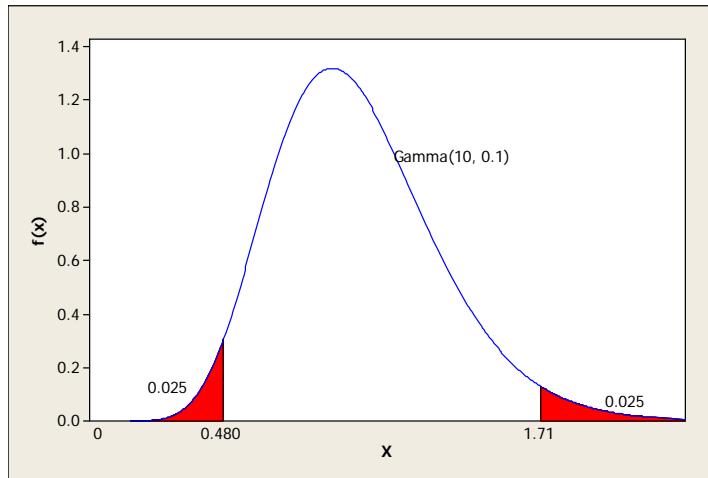
When sample sizes are small, the standard normal approximations are typically unjustifiable. Under these circumstances, one must obtain the appropriate sampling distribution for the particular problem at hand, and then use it to determine the interval estimate. Let us illustrate this concept for samples drawn from an exponential population, using the data of Example 14.3.

#### Example 14.14: INTERVAL ESTIMATES FOR EXPONENTIAL DISTRIBUTION MEANS

From the data on waiting time (in days) until the occurrence of a recordable safety incident in a certain company's manufacturing site given in Example 14.3, obtain a 95% confidence interval for this exponential random variable's unknown parameter  $\beta$ , first for the first year data set  $S_1$  and then for the second year data set  $S_2$ . Compare these interval estimates.

**Solution:**

The point estimate  $\hat{\beta}$  for the first year data was obtained from the sample average in Example 14.3 as 30.1 (this is also the MLE). Now, if  $X_i \sim E(\beta)$ , then from results presented in Chapter 6, or more specifically, from Eq (9.39) in Chapter 9, we know that the random variable



**FIGURE 14.4:** Sampling distribution with two-sided tail area probabilities of 0.025 for  $\bar{X}/\beta$ , based on a sample of size  $n = 10$  from an exponential population

$\bar{X}$ , defined as

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (14.125)$$

has the Gamma distribution  $\gamma(n, \beta/n)$ . However, note that this pdf depends on the unknown parameter  $\beta$  and can therefore not be used, as is, to make probabilistic statements. On the other hand, by scaling  $\bar{X}$  with  $\beta$ , we see that

$$\frac{1}{\beta} \bar{X} \sim \gamma(n, 1/n) \quad (14.126)$$

a pdf that now depends only on the sample size,  $n$ . (This is directly analogous to the  $t$ -distribution which depends only on the degrees of freedom,  $(n - 1)$ ).

And now for the specific case with  $n = 10$ , we obtain from the  $\text{Gamma}(10, 0.1)$  distribution the following:

$$P\left(0.48 < \frac{\bar{X}}{\beta} < 1.71\right) = 0.95 \quad (14.127)$$

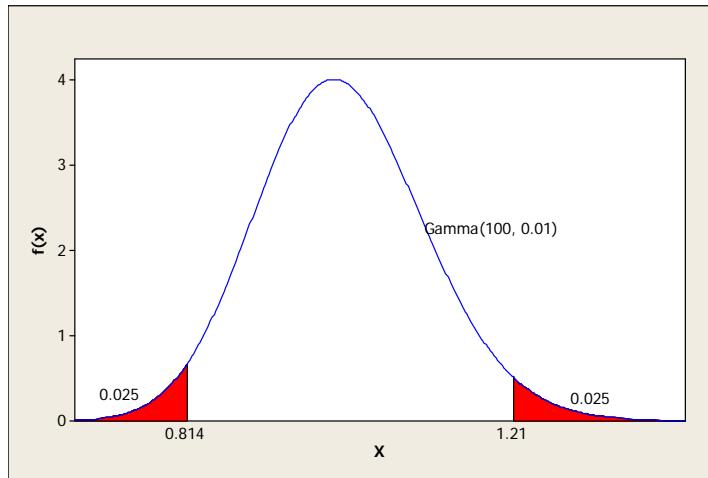
(see Fig 14.4) where the values for the interval boundaries are obtained from MINITAB using the inverse cumulative probability feature.

For the specific case of the first year data with  $\bar{x} = 30.1$ , the expression in Eq (15.144) may then be rearranged to yield the 95% confidence interval:

$$17.6 < \beta_1 < 62.71 \quad (14.128)$$

and for the second year data, with  $\bar{x} = 32.9$ ,

$$19.24 < \beta_2 < 68.54 \quad (14.129)$$



**FIGURE 14.5:** Sampling distribution with two-sided tail area probabilities of 0.025 for  $\bar{X}/\beta$ , based on a *larger* sample of size  $n = 100$  from an exponential population

First, note the asymmetry in these intervals in relation to the respective values for the point estimates,  $\bar{X}$ : this should not come as a surprise, since the Gamma distribution is skewed to the right. Next, observe that these intervals are quite wide; this is primarily due to the relatively small sample size of 10. Finally, observe that the two intervals overlap considerably, suggesting that the two estimates may not be different at all; the difference of 2.8 from year 1 to year 2 is more likely due to random variation than to any actual systemic improvement.

To use this last example to illustrate the applicability of the standard normal approximation for large sample sizes, let us consider that the sample averages  $\bar{x}_1 = 30.1$  and  $\bar{x}_2 = 32.9$  were actually obtained from a sample size of 100. Under these conditions, the sampling distribution for  $\bar{X}/\beta$  will be  $\text{Gamma}(100, 0.01)$ , and the values of the random variable that will yield two-sided tail area probabilities of 0.025 are obtained from MINITAB as 0.814 and 1.21 (see Fig 14.5). For the first year data with  $\bar{x} = 30.1$ , we are then able obtain the 95% confidence interval:

$$\frac{\bar{X}}{1.21} < \beta_1 < \frac{\bar{X}}{0.814} = 24.88 < \beta_1 < 36.98 \quad (14.130)$$

and, similarly for the second year data, with  $\bar{x} = 32.9$ ,

$$27.19 < \beta_2 < 40.42 \quad (14.131)$$

Not surprisingly, these intervals are considerably tighter than the corresponding ones obtained for the smaller sample size  $n = 10$ .

We now wish to compare these intervals with ones obtained by invoking the approximate  $N(0, 1)$  distribution for means computed from large samples. For this, we need sample standard deviations, which we have not had any use for until now; these are obtained from the data sets  $S_1$  and  $S_2$  in Example 14.3 as  $s_1 = 23.17$ ;  $s_2 = 27.51$ . If we assume that these are reasonably close approximations to the true population standard deviations (which would be the case had we actually used a sample size of 100), then we obtain the approximate 95% confidence intervals as follows:

$$\beta_1 = 30.1 \pm 1.96 \frac{23.17}{\sqrt{100}} = 30.1 \pm 4.54 \quad (14.132)$$

$$\beta_2 = 32.9 \pm 1.96 \frac{22.51}{\sqrt{100}} = 32.9 \pm 5.39 \quad (14.133)$$

which, when written in the same form as in Eqs (14.130) and (14.131), yields

$$25.56 < \beta_1 < 34.54 \quad (14.134)$$

$$27.51 < \beta_2 < 38.29 \quad (14.135)$$

We see that the approximation is in fact quite good.

## 14.6 Bayesian Estimation

### 14.6.1 Background

In all our discussion until now, the unknown population parameters,  $\theta$ , have been considered as *fixed*, deterministic constants whose values we have sought to determine solely on the basis of information contained in sample data drawn from the population in question. No *a-priori* knowledge or information about  $\theta$  is considered or assumed. And the results obtained from the estimation techniques presented thus far have either been the best single point values for each parameter (point estimates), or else appropriate intervals that we expect to contain each parameter, with a pre-specified probability, typically 0.95 (interval estimates).

There is an alternative, fundamentally different approach, with the following defining characteristics:

1. The unknown parameters are considered as random variables,  $\Theta$ , with  $\theta$  considered as a specific value of the random vector whose pdf is  $f(\theta)$ ;
2. Instead of providing point estimates along with a probabilistic interval, the objective is to obtain the full pdf, from which all sorts of probabilistic statements can be made;

3. Any available prior information about the random vector  $\Theta$  and its pdf, can and should be used in conjunction with sample data in providing parameter estimates.

This approach is known as “Bayesian” Estimation. Its basis is the fundamental relationship between joint, conditional and marginal pdfs, which we may recall from Chapter 4 as:

$$f(x|y) = \frac{f(x,y)}{f(y)} \quad (14.136)$$

from which one obtains the following important result

$$f(x,y) = f(x|y)f(y) = f(y|x)f(x) \quad (14.137)$$

that is used to reverse conditional probabilities, since upon rearrangement Eq (14.137) becomes:

$$f(y|x) = \frac{f(x|y)f(y)}{f(x)} \quad (14.138)$$

This expression is known as Bayes’ Theorem; it is attributed to the Revd Thomas Bayes (1702–1761), a Presbyterian minister and something of an amateur mathematician in the original sense of the word “amateur.” The theorem that bears his name appeared in his now-famous, posthumously-published, paper; but the subject of that paper was in fact just a special case of the more general version later proved by Laplace (1749–1827).

#### 14.6.2 Basic Concept

Consider a random sample  $X_1, X_2, \dots, X_n$  from a population with pdf  $f(x; \boldsymbol{\theta})$  and unknown parameters  $\boldsymbol{\theta}$ ; we know that the joint pdf is given by:

$$f(x_1, x_2, \dots, x_n | \boldsymbol{\theta}) = f(x_1; \boldsymbol{\theta})f(x_2; \boldsymbol{\theta}) \cdots f(x_n; \boldsymbol{\theta}) \quad (14.139)$$

This is the conditional pdf of the data conditioned on  $\boldsymbol{\theta}$ ; for any given value of  $\boldsymbol{\theta}$ , this expression provides the probability of jointly observing the data  $\{x_1, x_2, \dots, x_n\}$  in the discrete case. For the continuous case, it is the density function to be used in computing the appropriate probabilities. (Recall that we earlier referred to this same expression as the likelihood function  $L(\boldsymbol{\theta})$  in Eq 14.40.)

Now, if  $\Theta$  is considered a random variable for which  $\boldsymbol{\theta}$  is just one possible realization, then in trying to determine  $\boldsymbol{\theta}$ , what we desire is the conditional probability of  $\Theta$  given the data; i.e.  $f(\boldsymbol{\theta}|x_1, x_2, \dots, x_n)$ , the reverse of Eq (14.139). This is obtained by invoking Bayes’ Theorem,

$$f(\boldsymbol{\theta}|x_1, x_2, \dots, x_n) = \frac{f(x_1, x_2, \dots, x_n | \boldsymbol{\theta})f(\boldsymbol{\theta})}{f(x_1, x_2, \dots, x_n)} \quad (14.140)$$

where

$$\begin{aligned} f(x_1, x_2, \dots, x_n | \boldsymbol{\theta}) &\equiv \text{the sampling distribution} \\ f(\boldsymbol{\theta}) &\equiv \text{the prior distribution of } \boldsymbol{\theta} \\ f(x_1, x_2, \dots, x_n) &\equiv \text{the marginal distribution of the data.} \end{aligned}$$

$f(\boldsymbol{\theta})$  is the *marginal* distribution of  $\boldsymbol{\Theta}$  without considering the data, a distribution defined *a-priori*, before acquiring data (and independent of the current data set). As a result of the convention of referring to  $f(\boldsymbol{\theta})$  as the *prior* distribution for  $\boldsymbol{\Theta}$ ,  $f(\boldsymbol{\theta}|x_1, x_2, \dots, x_n)$  is referred to as the posterior (or *a-posteriori*) distribution of  $\boldsymbol{\Theta}$  because it is obtained after acquiring data (i.e. conditioned upon the observed data).

Now, because  $x_1, x_2, \dots, x_n$  constitutes an observed data set with known values, any function of such known quantities is itself known. As such,  $f(x_1, x_2, \dots, x_n)$ , regardless of its actual functional form, is a known constant once the observation  $x_1, x_2, \dots, x_n$  is given. Thus, we may rewrite Eq (14.140) as

$$\begin{aligned} f(\boldsymbol{\theta}|x_1, x_2, \dots, x_n) &= Cf(\boldsymbol{\theta})f(x_1, x_2, \dots, x_n | \boldsymbol{\theta}) & (14.141) \\ \text{POSTERIOR} &\propto \text{PRIOR} \times \text{SAMPLING} \end{aligned}$$

Thus, through Eq (14.141), the posterior pdf of  $\boldsymbol{\Theta}$  combines prior information about  $\boldsymbol{\Theta}$  available as  $f(\boldsymbol{\theta})$  with information from sample data available as  $f(x_1, x_2, \dots, x_n | \boldsymbol{\theta})$  (more compactly,  $f(\mathbf{x}|\boldsymbol{\theta})$ ).

#### 14.6.3 Bayesian Estimation Results

The primary result of Bayesian estimation is  $f(\boldsymbol{\theta}|\mathbf{x})$ , the posterior pdf for  $\boldsymbol{\Theta}$  conditioned upon the observed data vector  $\mathbf{x}$ ; no point or interval estimates are given directly. However, since  $f(\boldsymbol{\theta}|\mathbf{x})$  is a full pdf, both point and interval estimates are easily obtained from it. For example, the mean, median, mode, or for that matter, any reasonable quantile of  $f(\boldsymbol{\theta}|\mathbf{x})$  can be used as a point estimate; and any interval,  $q_{1-\alpha/2} < \boldsymbol{\theta} < q_{\alpha/2}$ , encompassing an area of probability  $(1 - \alpha)$  can be used as an interval estimator. In particular,

1. The mean of the posterior pdf,

$$E[\boldsymbol{\Theta}|X_1, X_2, \dots, X_n] = \int_{\boldsymbol{\theta} \in \Theta} \boldsymbol{\theta} f(\boldsymbol{\theta}|x_1, x_2, \dots, x_n) d\boldsymbol{\theta} \quad (14.142)$$

is called the *Bayes' estimator*.

2. The mode of  $f(\boldsymbol{\theta}|x_1, x_2, \dots, x_n)$  is called the *maximum a-posteriori* (MAP) estimator.

The typical procedure for carrying out Bayesian estimation may now be summarized as follows: Given a random variable,  $X$ , with pdf  $f(x; \boldsymbol{\theta})$ ,

1. Begin by specifying a prior distribution,  $f(\boldsymbol{\theta})$ , a summary of prior knowledge about the unknown parameters  $\boldsymbol{\theta}$ ;
2. Obtain sample data in the form of a random sample  $X_1, X_2, \dots, X_n$ , and hence the sampling distribution (the joint pdf for these  $n$  independent random variables with identical pdfs  $f(x_i; \boldsymbol{\theta})$ );
3. From Eq (14.141) obtain the posterior pdf,  $f(\boldsymbol{\theta}|x_1, x_2, \dots, x_n)$ ; (if needed, determine  $C$  such that  $f(\boldsymbol{\theta}|x_1, x_2, \dots, x_n)$  is a true pdf, i.e.

$$\frac{1}{C} = \int_{\boldsymbol{\theta} \in \Theta} f(\boldsymbol{\theta}) f(x_1, x_2, \dots, x_n | \boldsymbol{\theta}) d\boldsymbol{\theta} \quad (14.143)$$

4. If point estimates are required, obtain these from the posterior distribution.

#### 14.6.4 A Simple Illustration

We consider the problem of determining the Binomial/Bernoulli “probability of success” parameter,  $p$ .

##### Data

Upon performing the required experiment, suppose that the result (after  $n$  trials) is  $X_i, i = 1, 2, \dots, n$ , a random sample of Bernoulli variables for which, as usual,  $X_i = 1$  for “success” and 0 for “failure.” Suppose that the total number of successes is determined to be  $X$  (i.e.  $X = \sum_{i=1}^n X_i$ ). We know that the sampling distribution of  $X$  is the binomial pdf:

$$f(x|\theta) = \binom{n}{X} \theta^X (1-\theta)^{n-X} \quad (14.144)$$

(We could also treat this as a random sample from a Bernoulli population; the resulting joint pdf will be the same as in Eq (14.144), up to the multiplicative constant  $\binom{n}{X}$ .)

##### The Prior Distribution

The signature feature of Bayesian estimation is the prior distribution and its usage. There are several ways to decide on the prior distribution for  $\theta$ , the Binomial/Bernoulli parameter, and we will consider two.

**CASE I:** We know that  $\theta$  can only take values between 0 and 1, restricting its range to the region  $[0,1]$ . If there is no additional *a-priori* information available about the parameter, then we could invoke the maximum entropy result of Chapter 10 to determine that the best prior distribution under these circumstances is the uniform distribution on the unit interval, i.e.:

$$f(\theta) = 1; 0 < \theta < 1 \quad (14.145)$$

**CASE II:** In addition to the known range, there is also prior information, for example, from a similar process for which on average, the probability of success is 0.4, with a variability captured by a variance of 1/25. Under these circumstances, again the maximum entropy results of Chapter 10 suggest a beta distribution with parameters  $(\alpha, \beta)$  determined from the prescribed mean and variance. From the expressions for the mean and variance of the  $Beta(\alpha, \beta)$  random variable given in Chapter 9, we are able to solve the two equations simultaneously to obtain

$$\alpha = 2; \beta = 3 \quad (14.146)$$

the prescribed prior pdf is therefore

$$f(\theta) = 12\theta(1 - \theta)^2 \quad (14.147)$$

### The Posterior Distribution and Point Estimates

We are now in a position to obtain posterior distributions for each case. For CASE I, the posterior distribution is obtained from Eq (14.141) as:

$$f(\theta|x) = C \times 1 \times \binom{n}{x} \theta^x (1 - \theta)^{n-x} \quad (14.148)$$

where the specific observation  $X = x$  has been introduced. To complete the determination of the posterior pdf, we need to determine the constant  $C$ . There are several ways to do this: by integration, as noted earlier (see Eq (14.143)), or by inspection — by which we mean that, as a function of  $\theta$ , Eq (14.148) looks like a Beta pdf, a fact that can be exploited as follows: From the exponents of  $\theta$  and of  $(1 - \theta)$ , this posterior pdf looks like the pdf of a  $Beta(\alpha, \beta)$  random variable with  $\alpha - 1 = x$ , and  $\beta - 1 = n - x$ , implying that

$$\alpha = x + 1; \beta = n - x + 1 \quad (14.149)$$

This being the case, the multiplying constants in Eq (14.148) must therefore be  $\Gamma(\alpha + \beta)/\Gamma(\alpha)\Gamma(\beta)$ . And because  $n$  and  $x$  are both integers in this problem, we are able to use the factorial representation of the Gamma function (see Eq (9.25)) to obtain the complete posterior pdf as:

$$f(\theta|x) = \frac{(n+1)!}{x!(n-x)!} \theta^x (1 - \theta)^{n-x} \quad (14.150)$$

It is left as an exercise to the reader to establish that, upon completing the integration in

$$C \int_0^1 \binom{n}{x} \theta^x (1 - \theta)^{n-x} d\theta = 1 \quad (14.151)$$

and solving for  $C$ , the result is

$$C = n + 1 \quad (14.152)$$

so that

$$C \binom{n}{x} = \frac{(n+1)!}{x!(n-x)!} \quad (14.153)$$

as implied in Eq (14.150).

We may now choose to leave the result as the pdf in Eq (14.150) and use it to make probabilistic statements about the parameter; alternatively, we can determine point estimates from it. For example, the Bayes's estimate,  $\hat{\theta}_B$ , defined as  $E[\theta|X]$  is obtained from Eq (14.150) as:

$$\hat{\theta}_B = \frac{x+1}{n+2} \quad (14.154)$$

where the result is obtained immediately by virtue of the posterior pdf being a Beta distribution with parameters given in Eq (14.149); or else (the hard way!) by computing the required expectation via direct integration. On the other hand, the MAP estimate,  $\hat{\theta}^*$ , is obtained by finding the maximum (mode) of the posterior pdf: from Eq (14.150) one obtains this as:

$$\hat{\theta}^* = \frac{x}{n}, \quad (14.155)$$

the same as the MLE.

For the sake of completeness, let us now suppose that after performing an actual series of experiments on a sample of size  $n = 10$ , one obtains 3 successes: this specific experimental result generates the following point estimates for the Binomial/Bernoulli probability of success:

$$\hat{\theta}_B = 4/12 = 0.33; \hat{\theta}^* = 3/10 = 0.3 \quad (14.156)$$

Thus, using a uniform prior distribution for  $f(\theta)$ , produces a Bayes estimate, estimate of  $1/3$ , compared to MAP estimate of  $0.3$  (which coincides with the standard MLE estimate).

Note that in CASE I, the prior distribution is somewhat “uninformative” and non-subjective, in the sense that it showed no preference for any value of  $\theta$  *à-priori*. Note that since  $x/n$  is known to be an unbiased estimate for  $\theta$ , then  $\hat{\theta}_B$  in Eq (14.154) is biased. However, it can be shown, (see Exercises 14.8, 14.34 and 14.33) that the variance of  $\hat{\theta}_B$  is always less than that of the unbiased MLE,  $\hat{\theta} = x/n$ . Thus, the Bayes estimate may be biased, but it is more efficient.

CASE II is different. The possible values of  $\theta$  are not assigned equal *à-priori* probability; the *à-priori* probability specified in Eq (14.147) definitely favors some values over others. We shall return shortly to the obvious issue

of subjectivity that this approach raises. For now, with the pdf given in Eq (14.147), the resulting posterior pdf is:

$$f(\theta|x) = C \times 12 \times \binom{n}{x} \theta^{x+1} (1-\theta)^{n-x+2} \quad (14.157)$$

It is straightforward to establish (see Exercise 14.33) that the final posterior pdf is given by

$$f(\theta|x) = \frac{(n+4)!}{(x+1)!(n-x+2)!} \theta^{x+1} (1-\theta)^{n-x+2} \quad (14.158)$$

so that the Bayes estimate and the MAP estimate in this case are given respectively as:

$$\hat{\theta}_B = \frac{x+2}{n+5} \quad (14.159)$$

and

$$\hat{\theta}^* = \frac{x+1}{n+3} \quad (14.160)$$

Again, with the specific experimental outcome of 3 successes in  $n = 10$  trials, we obtain the following as the CASE II estimates

$$\hat{\theta}_B = 5/15 = 0.33; \hat{\theta}^* = 4/13 = 0.31 \quad (14.161)$$

It is important to note that as different as CASE I and CASE II conditions are, and as different as the criteria for determining the Bayes estimates and the MAP estimates are, the results are all quite similar. Still, as we noted in Case I, so it is in this case that the estimates,  $\hat{\theta}_B$  and  $\hat{\theta}^*$ , are biased, but their variances can be shown to be smaller than that of the MLE estimate.

#### 14.6.5 Discussion

##### The Bayesian Controversy: Subjectivity

Bayesian estimation is considered controversial in many scientific circles. The primary reason is that, in contrast to standard (or “frequentist”) estimation, which is based entirely on data alone, Bayesian estimation combines prior information with data to obtain parameter estimates. Since the prior distribution is mostly a “subjective” description of the variability in the unknown parameter vector  $\theta$  *before* data is acquired, the argument against Bayesian estimation is that it introduces subjectivity into data analysis.

In many other circles, however, the very fact that Bayesian estimation provides a systematic methodology for incorporating prior knowledge into data analysis is what makes it quite attractive. Consider, for example, a study of the reliability of a new generation of power distribution networks. Such studies usually build on results from previous studies on earlier generations, which themselves were follow-ups to earlier studies. While each new generation of

networks have their own characteristics and properties, ignoring earlier studies of previous generations as if they did not exist is not considered good engineering practice. Whatever relevant prior information is available from previous studies should be incorporated into the current analysis. Many areas of theoretical and applied sciences (including engineering) advance predominantly by building on prior knowledge, not by ignoring available prior information.

Still, the possibility remains that the subjectivity introduced into data analysis by the choice of the prior distribution,  $f(\theta)$ , could dominate the objective information contained in the data. The counter-argument is that the Bayesian approach is actually completely transparent in how it distinctly separates out each component of the entire data analysis process — what is subjective and what is objective — making it possible to assess, in an objective manner, the influence of the prior information on the final result. It also allows room for adaptation, in light of additional objective information.

### Recursive Bayesian Estimation

It is in the latter sense noted above that the Bayesian approach provides its most compelling advantage: recursive estimation. Consider a case that is all too common in the chemical industry in which the value of a process variable, say viscosity of a polymer material, is to be determined experimentally. The measured value of the process variable is subject to random variation by virtue of the measurement device characteristics, but also intrinsically. In particular, the true value of such a variable changes dynamically (i.e. from one time instant to the next, the value will change because of dynamic operating conditions). If the objective is to estimate the current value of such a variable, the “frequentist” approach as discussed in the earlier parts of this chapter, is to obtain a random sample of size  $n$  from which the true value will be estimated. Unfortunately, because of dynamics, only a single value is obtainable at any point in time,  $t_k$ , say  $x(t_k)$ ; at the next sampling point, the observed value,  $x(t_{k+1})$  is, technically speaking, not the same as the previous value, and in any event, the two can hardly be considered as independent of each other. There is no realistic frequentist solution to this problem. However, by postulating a prior distribution, one can obtain a posterior distribution on the basis of a single data point; the resulting posterior distribution, which now incorporates the information contained in the just-acquired data, can now be used as the prior distribution for the next round. In this recursive strategy, the admittedly subjective prior employed as an “initial condition” is ultimately washed out of the system with progressive addition of objective, but time-dependent data. A discussion of this type of problem is included in the application case studies of Chapter 20.

### Choosing Prior Distributions

The reader may have noticed that the choice of a Beta distribution as the prior distribution,  $f(\theta)$ , for the Binomial/Bernoulli probability of success,  $p$ , is

particularly “appropriate.” Not only is the Beta random variable conveniently scaled between 0 and 1 (just like  $p$ ), the functional form of the Beta pdf is perfectly paired with the Binomial pdf, when viewed from the perspective of the unknown parameter,  $p$ . The two pdfs are repeated here for ease of comparison:

$$f(x|\theta) = \binom{n}{X} \theta^X (1-\theta)^{n-X} \quad (14.162)$$

$$f(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \quad (14.163)$$

where, even though the first is a function of the random variable,  $X$ , and the second is a function of the parameter,  $\theta$ , the two pdfs are seen to have what is called a “conjugate” structure: multiplying one by the other results in a posterior pdf where the “conjugate” structure is preserved. The Beta pdf is therefore said to provide a conjugate prior for the Binomial sampling distribution. The advantage of employing conjugate priors is therefore clear: it simplifies the computational work involved in determining the posterior distribution.

For the Poisson  $\mathcal{P}(\lambda)$  random variable with unknown parameter  $\lambda = \theta$ , the conjugate prior is the Gamma distribution. Arranged side-by-side, the nature of the conjugate structure becomes obvious:

$$f(x|\theta) = \frac{e^{-\theta}\theta^x}{x!}; x = 0, 1, 2, \dots \quad (14.164)$$

$$f(\theta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} e^{-\theta/\beta} \theta^{\alpha-1}; 0 < \theta < \infty \quad (14.165)$$

(This particular prior will be used in a case study in Chapter 20.)

There are a few more such conjugate sampling and prior distribution pairs, for example, the normal sampling distribution and the normal prior distribution; the exponential sampling distribution (with  $\theta = 1/\beta$ ) and the gamma prior distribution. Some of these are listed in Table 14.2.

We conclude by noting that while conjugate priors provide sampling distribution pairings that simplify analytical determination of posterior distributions, in fact, it is not necessary to seek conjugate priors for all Bayesian estimation problems. The most appropriate prior distribution should be selected even if it does not form a “conjugate” pair with the sampling distribution.

### Computational Issues

In more general cases, where any appropriate prior distribution is combined with the sampling distribution in question, determining the resulting posterior distribution is not always a trivial matter because this exercise usually involves computing multi-dimensional integrals. Under such general conditions, it is not always possible to obtain explicit forms for the posterior

distributions. In many cases, the only option is to obtain the required posterior distributions (as well as point estimates) numerically. Until recently, the computational burden of numerically determining posterior distributions for practical problems constituted a considerable obstacle to the application of Bayesian techniques in estimation. With the introduction of the Markov Chain Monte Carlo (MCMC)<sup>1</sup> techniques, however, this computational issue has essentially been resolved. There are now commercial software packages for carrying out such numerical computations quite efficiently.

---

## 14.7 Summary and Conclusions

Our study of statistical inference began in this chapter with estimation—the process by which unknown population parameters are determined from limited sample information—building directly on the foundation of sampling theory laid down in Chapter 13. We were primarily concerned with techniques for obtaining point estimates and how to quantify their precision, leading naturally to interval estimates. The method of moments technique might have appeared a bit *ad-hoc* (because it does not produce unique estimates), but it is quite straightforward, intuitive and quite useful in providing initial estimates that may be subsequently refined, if necessary. And, in any event, there are many cases where these estimates coincide precisely with the more systematic maximum likelihood estimates. On the other hand, most readers will probably admit to sensing something of a “much ado about nothing” air surrounding the method of maximum likelihood. For example, why go through all the calculus and algebra only to discover the obvious—that the sample mean is the MLE for the population mean? It is instructive, however, to keep in mind that such simple closed form MLE solutions exist only in a handful of cases. No such solution exists for the gamma distribution parameters, for example, and definitely not for the Weibull distribution parameters. In a sense, therefore, it ought to be taken as a reassuring sign that in obvious cases, the maximum likelihood principle produces such intuitively obvious results; this should give the reader confidence that in cases where the results must be computed numerically, these results can also be trusted.

The nature and characteristics of the distribution of random samples—especially their variances—made it difficult to present any general results about interval estimates for anything other beyond normal populations. Our brief discussion of how to obtain such interval estimates for non-Gaussian populations (when samples sizes are small) should be understood as illustrations

---

<sup>1</sup>Gilks W.R., Richardson S. and Spiegelhalter D.J. *Markov Chain Monte Carlo in Practice*, Chapman & Hall/CRC, 1996.

of what is possible when Gaussian approximations are invalid; generalized discussions are practically impossible.

The discussion of criteria for selecting estimators at the beginning of the chapter also may have seemed somewhat “obvious” and superfluous, until we encountered Bayesian estimation in the final section of the chapter and the issue of bias and efficiency became important. If the reader had been wondering why anyone would ever consider using anything but an *unbiased* estimator, or questioned what practical implications the concept of “efficiency” could possibly have, Bayesian estimation put both issues in context simultaneously. Sometimes, especially when sample sizes are small, it may make more sense to opt for a biased estimator with a smaller variance. Such is the case with Bayesian estimation. Admittedly, the discussion of Bayesian estimation in this chapter was rather brief, but that does not make this estimation technique any less important. We have tried to make up for this by providing several exercises and a few illustrative application problems designed to expand on the brief coverage.

Finally, we note that many of the seeds of hypothesis testing have been sown already in this chapter; we shall see these fully developed in the next chapter when we bring the discussion of statistical inference to a conclusion. For now, a summary of the key results of this chapter is summarized in Table 14.1 along with some information about Bayesian estimation in Table 14.2.

---

## REVIEW QUESTIONS

1. The objective of this chapter is to provide answers to what sorts of questions?
2. What is “estimation”?
3. What are the two aspects of estimation discussed in this chapter?
4. What is an estimator?
5. What is a point estimate and how is it different from an estimator?
6. What is an interval estimator?
7. What is an interval estimate and how is it different from an interval estimator?
8. What is the mathematical definition of an unbiased estimator?
9. What makes unbiasedness an intuitively appealing criterion for selecting estimators?

- 10.** Mathematically, what does it mean that one estimator is more efficient than another?
- 11.** What is the mathematical definition of a consistent sequence of estimators?
- 12.** What is the basic principle behind the method of moments technique for obtaining point estimates?
- 13.** Are method of moments estimators unique?
- 14.** What is the likelihood function and how is it differentiated from the joint pdf of a random sample?
- 15.** What is the log-likelihood function? Why is it often used in place of the likelihood function in obtaining point estimates?
- 16.** Are maximum likelihood estimates always unbiased?
- 17.** What is the invariance property of maximum likelihood estimators?
- 18.** What are the asymptotic properties of maximum likelihood estimators?
- 19.** What is needed in order to quantify the precision of point estimates?
- 20.** What are the two main components of an interval estimate?
- 21.** What is the general procedure for determining interval estimates for the mean of a normal population with  $\sigma$  known?
- 22.** What is the difference between interval estimates for the mean of a normal population when  $\sigma$  is known and when  $\sigma$  is unknown?
- 23.** What probability distribution is used to obtain interval estimates for the variance of a normal population?
- 24.** Why is the confidence interval around the point estimate of a normal population variance not symmetric?
- 25.** How can interval estimates of the difference between two normal population means be used to assess the equality of these means?
- 26.** How does one obtain interval estimates for parameters from other non-Gaussian populations when samples sizes are large and when they are small?
- 27.** What are the distinguishing characteristics of Bayesian estimation?
- 28.** What is a prior distribution,  $f(\theta)$ , and what role does it play in Bayesian estimation?

- 29.** Apart from the prior distribution, what other types of probability distributions are involved in Bayesian estimation, and how are they related?
- 30.** What is the primary result of Bayesian estimation?
- 31.** What is the Bayes estimator? What is the maximum *a-posteriori* estimator?
- 32.** What are some of the controversial aspects of Bayesian estimation?
- 33.** What is recursive Bayesian estimation?
- 34.** What is a “conjugate” prior distribution?
- 35.** What are some of the computational issues involved in the practical application of Bayesian estimation, and how have they been resolved?

## EXERCISES

### Section 14.1

**14.1** Given a random sample  $X_1, X_2, \dots, X_n$ , from a Gaussian  $N(\mu, \sigma^2)$  population with unknown parameters, it is desired to use the sample mean,  $\bar{X}$ , and the sample variance,  $S^2$ , to determine point estimates of the unknown parameters;  $\bar{X} \pm 2S^2/\sqrt{n}$  is to be used to determine an interval estimate. The following data set was obtained for this purpose:

$$\{9.37, 8.86, 11.49, 9.57, 9.15, 9.10, 10.26, 9.87\}$$

- (i) In terms of the random sample, what is the estimator for  $\mu$ ; and in terms of the supplied data, what is the point estimate for  $\mu$ ?
- (ii) What are the boundary estimators,  $U_L$  and  $U_R$ , such that  $(U_L, U_R)$  is an interval estimator for  $\mu$ ? What is the interval estimate?

**14.2** Refer to Exercise 14.1.

- (i) In terms of the random sample, what is the estimator for  $\sigma^2$ , and what is the point estimate?
- (ii) Given the boundary estimators for  $\sigma^2$  as:

$$U_L = \frac{(n-1)S^2}{C_L}; U_R = \frac{(n-1)S^2}{C_R}$$

where  $C_L = 19.0$  and  $C_R = 2.7$ , obtain an interval estimate for  $\sigma^2$ .

**14.3** Consider a random sample  $X_1, X_2, \dots, X_n$  from a population with unknown

mean,  $\mu$ , and variance,  $\sigma^2$ , and the following estimators,

$$\begin{aligned} M_1 &= \frac{1}{n} \sum_{i=1}^n X_i; \\ M_2 &= \frac{1}{n} \sum_{i=1}^n X_i^2 \end{aligned}$$

If the following relationship holds:

$$E(X^2) = Var(X) + [E(X)]^2$$

where  $Var(X)$  is the variance of the random variable, determine the estimator for  $Var(X)$  in terms of  $M_1$  and  $M_2$ . Determine the point estimates,  $m_1, m_2$  and  $s^2$  respectively for  $M_1, M_2$  and  $Var(X)$  from the following sample data:

$$\{10.09, 15.73, 15.04, 5.55, 18.04, 17.95, 12.55, 9.66\}$$

**14.4** Consider a random sample  $X_1, X_2, \dots, X_n$  from a population with unknown mean,  $\mu$ , and variance,  $\sigma^2$ ; define the following estimator

$$\bar{X}_g = \left( \prod_{i=1}^n X_i \right)^{1/n}$$

- (i) In terms of the supplied information, what is the estimator for  $A = \ln \bar{X}_g$  and for  $\Xi = (\bar{X}_g)^n$ ?
- (ii) Determine point estimates for  $\bar{X}_g$ ,  $A$ ,  $e^A$  and for  $\bar{X}$ , the sample mean, from the following sample data:

$$\{7.91, 5.92, 4.53, 33.26, 24.13, 5.42, 16.96, 3.93\}$$

### Section 14.2

**14.5** Consider a random sample  $X_1, X_2, \dots, X_n$ ;

- (i) If the sample is from a Lognormal population, i.e.  $X \sim \mathcal{L}(\alpha, \beta)$ , so that

$$E(\ln X) = \alpha$$

and if the sample geometric mean is defined as

$$\bar{X}_g = \left( \prod_{i=1}^n X_i \right)^{1/n}$$

show that  $\ln \bar{X}_g$ , not  $X_g$ , is unbiased for  $\alpha$ .

- (ii) If the sample is from an exponential population defined in terms of  $\eta$ , the *rate* of occurrence of the underlying Poisson events, i.e., the pdf is given as:

$$f(x) = \eta e^{-\eta x}; 0 < x < \infty$$

and if the sample mean is defined as

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

show that this estimator is unbiased for  $1/\eta$ , but the estimator  $1/\bar{X}$  is *not* unbiased for  $\eta$ . (You do not have to compute the expectation of  $1/\bar{X}$ ).

**14.6** Given a random sample  $X_1, X_2, \dots, X_n$  from a general population with unknown mean,  $\mu$ , and variance,  $\sigma^2$ , that the sample variance,

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

where  $\bar{X}$  is the sample mean, is unbiased for the population variance  $\sigma^2$ , regardless of the underlying population.

**14.7** Consider the estimator  $\hat{\Theta}$ , with mean  $\mu(\hat{\Theta})$  and variance  $\sigma^2(\hat{\Theta})$ , proposed as an estimator for the unknown population parameter,  $\theta$ . If  $\hat{\Theta}$  is a *biased* estimator i.e.,  $E(\hat{\Theta}) \neq \theta$ , define the bias  $B(\theta)$  as:

$$B(\theta) = E(\hat{\Theta}) - \theta = \mu(\hat{\Theta}) - \theta \quad (14.166)$$

and show that the mean squared estimation error, MSE, is given by

$$E[(\hat{\Theta} - \theta)^2] = \sigma^2(\hat{\Theta}) + (B(\theta))^2 \quad (14.167)$$

(Consider decomposing the estimation error into two components as follows:  $(\hat{\Theta} - \theta) = [\hat{\Theta} - \mu(\hat{\Theta})] + [\mu(\hat{\Theta}) - \theta]$ .)

**14.8** Given a random sample  $X_1, X_2, \dots, X_n$  from a Bernoulli population with unknown parameter  $\theta$ , then it is known first, that  $X = \sum_{i=1}^n X_i$  is a Binomial  $Bi(n, \theta)$  random variable; secondly, the estimator,

$$\hat{\Theta} = \frac{X}{n}$$

is unbiased for  $\theta$ . Consider a second estimator defined as:

$$\tilde{\Theta} = \frac{X+1}{n+2}$$

(i) Show that  $\tilde{\Theta}$  is biased for  $\theta$  and determine the bias  $B(\theta)$ , as defined in Eq (14.166) in Exercise 14.7.

(ii) Let  $\hat{V}$  and  $\tilde{V}$  represent the variances of the estimators  $\hat{\Theta}$  and  $\tilde{\Theta}$  respectively. Show that

$$\tilde{V} = \left( \frac{n}{n+2} \right)^2 \hat{V}$$

and hence establish that  $\tilde{V} < \hat{V}$ , so that the biased estimator  $\tilde{\Theta}$  is more efficient than  $\hat{\Theta}$ , the unbiased estimator, especially for small sample sizes,.

**14.9** Given a random sample  $X_1, X_2, \dots, X_n$  from a population with mean  $\mu$ , and variance  $\sigma^2$ , define two statistics as follows:

$$\begin{aligned} \bar{X} &= \frac{1}{n} \sum_{i=1}^n X_i \\ \tilde{X} &= \sum_{i=1}^n \omega_i X_i; \text{ with } \sum_{i=1}^n \omega_i = 1 \end{aligned}$$

It was shown in the text that both statistics are unbiased for  $\mu$ ; now show that  $\bar{X}$ , a special case of  $\tilde{X}$ , is the more efficient estimator of  $\mu$ .

**14.10** Given a random sample  $X_1, X_2, \dots, X_n$  from a general population with *finite* mean  $\mu$ , and finite variance  $\sigma^2$ , show that the sample mean,  $\bar{X} = (\sum_{i=1}^n X_i)/n$ , is consistent for  $\mu$  regardless of the underlying population. (Hint: Invoke the Central Limit Theorem.)

### Section 14.3

**14.11** Given a random sample,  $X_1, X_2, \dots, X_n$  from a Poisson  $\mathcal{P}(\lambda)$  population, obtain an estimate for the population parameter,  $\lambda$ , on the basis of the second moment estimator,

$$M_2 = \frac{1}{n} \sum_{i=1}^n X_i^2$$

and show that this estimate,  $\hat{\lambda}_2$ , is explicitly given by:

$$\hat{\lambda}_2 = \frac{1}{2} \left( \sqrt{4M_2 + 1} \right) - \frac{1}{2}$$

**14.12** Refer to Exercise 14.11 and consider the following sample of size  $n = 20$  from a Poisson  $\mathcal{P}(2.10)$  population.

|   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|
| 3 | 0 | 1 | 2 | 3 | 2 | 4 | 2 | 1 | 2 |
| 4 | 1 | 2 | 3 | 2 | 1 | 3 | 5 | 0 | 2 |

Let  $\hat{\lambda}_1$  represent the estimate obtained from the first moment, and  $\hat{\lambda}_2$  the estimate obtained from second moment, as in Exercise 14.11.

- (i) Determine specific numerical values for  $\hat{\lambda}_1$  and  $\hat{\lambda}_2$  from the data and compare them.
- (ii) Consider the following “weighted” estimate, a combination of these two estimates:

$$\hat{\lambda}_\omega = \omega \hat{\lambda}_1 + (1 - \omega) \hat{\lambda}_2$$

Determine the values of  $\hat{\lambda}_\omega$  for  $\omega = 0.1, 0.2, 0.3, \dots, 0.9$ . Plot the nine values as a function of  $\omega$  and compare these to the true value  $\lambda = 2.10$ .

**14.13** Given a random sample,  $X_1, X_2, \dots, X_n$  from a negative binomial  $NBi(k, p)$  population,

- (i) Obtain a method of moments estimate of  $\theta_1 = k$  and  $\theta_2 = p$  explicitly in terms of the first two moments,  $M_1$  and  $M_2$ .
- (ii) If  $k$  is specified (as it is in many cases), obtain two separate expressions for method of moments estimate for  $p$  based on  $M_1$  and  $M_2$ .
- (iii) When  $k$  is specified, obtain a maximum likelihood estimate (MLE) for  $p$  and compare it with the method of moments estimate obtained in (ii).

**14.14** (i) In terms of the first two moments  $M_1$  and  $M_2$ , obtain two separate method of moments estimators for the unknown parameter in the geometric  $G(p)$  distribution. Given the following data from a  $G(0.25)$  population, determine numerical values for the two different point estimates and compare them to the true population value.

|   |   |   |    |   |    |   |   |   |   |
|---|---|---|----|---|----|---|---|---|---|
| 1 | 6 | 2 | 2  | 9 | 14 | 2 | 1 | 2 | 1 |
| 1 | 1 | 5 | 14 | 2 | 2  | 6 | 3 | 1 | 1 |

(ii) Obtain the harmonic mean of the given data. As an estimate of the population parameter  $p = 0.25$ , how does this estimate compare with the two method of moments estimates for  $p$ ?

**14.15** In terms of the first two moments  $M_1$  and  $M_2$ , obtain two separate method of moments estimators for the unknown parameter in the exponential  $\mathcal{E}(\beta)$  distribution. From the following data sampled from an exponential  $\mathcal{E}(4)$  population, determine numerical values of the two different point estimates and compare them to the true population value.

|       |      |      |      |       |
|-------|------|------|------|-------|
| 6.99  | 2.84 | 0.41 | 3.75 | 2.16  |
| 0.52  | 0.67 | 2.72 | 5.22 | 16.65 |
| 10.36 | 1.66 | 3.26 | 1.78 | 1.31  |
| 5.75  | 0.12 | 6.51 | 4.05 | 1.52  |

**14.16** On the basis of the first two moments  $M_1$  and  $M_2$ , determine method of moments estimates for the two parameters in the Beta  $B(\alpha, \beta)$  distribution.

**14.17** Use the first two moments  $M_1$  and  $M_2$  to determine two separate estimators for the single Rayleigh  $\mathcal{R}(b)$  distribution parameter.

**14.18** On the basis of the first two moments  $M_1$  and  $M_2$ , determine method of moments estimates for the two parameters in the Gamma  $\gamma(\alpha, \beta)$  distribution.

**14.19** Show that the likelihood function for the binomial random variable given in Eq (14.66), i.e.,

$$L(p) = \binom{n}{X} p^X (1-p)^{n-X}$$

is maximized when  $\hat{p} = X/n$ , and hence establish the result stated in Eq (14.67).

**14.20** Let  $X_1, X_2, \dots, X_n$  be a random sample from a geometric population with unknown parameter  $\theta$ . Obtain the maximum likelihood estimate (MLE),  $\hat{\theta}$ . Show that

$$E(\hat{\theta}) \neq \theta$$

so that  $\hat{\theta}$  is in fact biased for  $\theta$ , but,

$$E\left(\frac{1}{\hat{\theta}}\right) = \frac{1}{\theta}$$

so that  $1/\hat{\theta}$  is unbiased for  $1/\theta$ .

**14.21** For the general negative binomial random variable  $NBi(k, p)$  with both parameters unknown, determine maximum likelihood estimates given the random sample,  $(k_i, X_i); i = 1, 2, \dots, n$ . When  $k$  is a fixed and known constant, so that the only unknown parameter is  $\theta = p$ , show that under these circumstances,  $\hat{\theta}$ , the MLE for  $\theta = p$ , is such that:

$$E(\hat{\theta}) \neq \theta$$

but,

$$E\left(\frac{1}{\hat{\theta}}\right) = \frac{1}{\theta}$$

**14.22** Given a random sample  $X_1, X_2, \dots, X_n$  from a Gamma  $\gamma(\alpha, \beta)$  population,  
 (i) Determine the MLE for the parameter  $\beta$  when  $\alpha$  is specified.

(ii) When  $\alpha$  is not specified, there is no closed form solution for the maximum likelihood estimates for the two unknown parameters. However, show, without solving the simultaneous maximum likelihood equations, that the method of moments estimates are different from the maximum likelihood estimates.

#### Sections 14.4 and 14.5

**14.23** The average of a sample of size  $n = 20$  from a normal population with unknown mean  $\mu$  and variance  $\sigma^2 = 5$  was obtained as  $\bar{x} = 74.4$ ; the sample variance was obtained as  $s^2 = 5.6$

- (i) Determine an interval  $(\bar{x} - w, \bar{x} + w)$  such that the probability that the true value of  $\mu$  lies in this interval is 0.90.
- (ii) Repeat (i) when the desired probability is 0.95
- (iii) Repeat (i) and (ii) when the population variance is not given.

**14.24** Refer to the data given in Table 1.2 in Chapter 1. Consider this a random sample of size  $n = 60$  drawn from a Poisson population with unknown parameter  $\lambda$ . The sample average was obtained in that chapter as  $\bar{x} = 1.02$  and is to be used as an estimate  $\hat{\lambda}$  of the unknown parameter. Determine the precision of this estimate by obtaining an interval such that

$$P(\bar{x} - w \leq \lambda \leq \bar{x} + w) = 0.95$$

Repeat for probability values 0.9 and 0.99.

**14.25** Let  $X_1, X_2, \dots, X_n$  be a random sample from a normal  $N(\mu, \sigma^2)$  population. With the sample variance  $S^2$  defined in Eq (14.99), (where  $\bar{X}$  is the sample average) we know that the statistic:

$$C = \frac{(n-1)S^2}{\sigma^2}$$

is a chi-squared distributed random variable; more specifically,  $C \sim \chi^2(n-1)$ . Show that the estimator  $S^2$  is unbiased for the population variance,  $\sigma^2$ , via the expected value of  $C$ .

**14.26** Given the following interval estimates for the mean of a random sample from a normal population with unknown mean but known variance,  $\sigma^2$ , determine the implied confidence levels:

- (i)  $\bar{X} \pm 1.645\sigma/\sqrt{n}$ ; (ii)  $\bar{X} \pm 2.575\sigma/\sqrt{n}$ ; (iii)  $\bar{X} \pm 3.000\sigma/\sqrt{n}$

**14.27** Let  $X_1, X_2, \dots, X_n$ , be a random sample from a normal population, with the distribution  $N(\mu_X, \sigma_X^2)$  which is independent of  $Y_1, Y_2, \dots, Y_m$ , another random sample from a different normal population, with the distribution  $N(\mu_Y, \sigma_Y^2)$ , where the sample sizes,  $n$  and  $m$  need not be equal (i.e.,  $n \neq m$ ).

- (i) Obtain the pdf for the random variable  $D = X - Y$ .

(ii) If  $\bar{X}$  is the MLE for  $\mu_X$ , and  $\bar{Y}$  is the MLE for  $\mu_Y$ , show that  $\bar{D}$ , defined as:

$$\bar{D} = \bar{X} - \bar{Y}$$

is the MLE for  $\delta_{XY} = \mu_X - \mu_Y$

**14.28** Let  $X_1, X_2, \dots, X_n$ , be a random sample from a normal population, with unknown mean, and variance  $\sigma^2 = 5$ . The population mean is to be estimated with the sample mean such that the 95% confidence interval will be  $\xi_L \leq \mu - \bar{X} < \xi_R$ , an interval of total width  $w = \xi_R - \xi_L$ .

- (i) Determine the sample size  $n$  required for  $w = 0.5$ .
- (ii) If the population variance doubles to  $\sigma^2 = 10$ , what value of  $n$  will be required to maintain the same width for the 95% confidence interval?
- (iii) If the population variance doubles but the sample size obtained in (i) is retained, what is the width of the 95% confidence interval?

**14.29** A random sample of size  $n = 100$  generated a sample mean,  $\bar{X} = 10.5$  and a sample variance of  $s^2 = 1.5$ . Estimate the unknown mean  $\mu$  with a 90% confidence interval, a 95% confidence interval, and a 99% confidence interval. Comment on how the width of the estimation intervals changes in relation to the desired confidence level and why this is entirely reasonable.

**14.30** An opinion poll based on a sample of 50 subjects estimated  $p$ , the proportion of the population in favor of the proposition, as 0.72.

- (i) Estimate the true proportion,  $\theta$ , with a 95% confidence interval. State any assumptions you may have to make in answering this question.
- (ii) If the true population proportion is suspected to be  $\theta = 0.8$ , and the estimate from an opinion poll is to be determined to within  $\pm 0.05$  with 95% confidence, how many people,  $n$ , should be sampled?
- (iii) If the proportion is to be estimated to within the same margin of  $\pm 0.05$ , but with 90% confidence, what is the value of  $n$  required? Comment on the effect that reducing the confidence level has on the sample size,  $n$  required to achieve the desired precision.

**14.31** The sample averages  $\bar{X} = 38.8$  and  $\bar{Y} = 42.4$  were obtained from random sample taken from two independent populations of respective sizes  $n_x = 120$  and  $n_y = 90$ . The corresponding sample standard deviations were obtained as  $s_x^2 = 20$ ;  $s_y^2 = 35$ . Determine a 95% confidence interval estimate for the difference  $\delta_{xy}$  using the difference between the sample averages. Does this interval include zero?

**14.32** Samples are to be taken from two different normal populations, one with variance  $\sigma_1^2 = 10$ , the other with a variance twice the magnitude of the first one. If the difference between the two population means is to be estimated to within  $\pm 2$ , with 95% confidence, determine the sample size required, assuming that  $n_1 = n_2 = n$ .

### Section 14.6

**14.33** In estimating an unknown binomial parameter, the posterior distribution arising from using a Beta  $B(2, 3)$  prior distribution was given in the text as Eq

(14.157), i.e.,

$$f(\theta|x) = C \times 12 \times \binom{n}{x} \theta^{x+1} (1-\theta)^{n-x+2}$$

(i) Show that in final form, with the constant  $C$  evaluated, this posterior distribution is:

$$f(\theta|x) = \frac{(n+4)!}{(x+1)!(n-x+2)!} \theta^{x+1} (1-\theta)^{n-x+2}$$

hence confirming Eq (14.158). (*Hint:* Exploit the structural similarity between this pdf and the Beta pdf.)

(ii) If  $x$  is the actual number of successes obtained in  $n$  trials, it is known that the estimate  $\hat{\theta} = x/n$  is unbiased for  $\theta$ . It was stated in the text that the Bayes and MAP estimates, are, respectively,

$$\hat{\theta}_B = \frac{x+2}{n+5}; \text{ and } \hat{\theta}^* = \frac{x+1}{n+3}$$

Show that these two estimates are *both* biased, but are both more efficient than  $\hat{\theta}$ . Which of the three is the most efficient?

**14.34** Let  $X_1, X_2, \dots, X_n$ , be a random sample from a Poisson distribution with unknown parameter,  $\theta$ .

(i) To estimate the unknown parameter with this sample, along with a Gamma  $\gamma(a, b)$  prior distribution for  $\theta$ , first show that the posterior distribution  $f(\theta|x)$  is a Gamma  $\gamma(a^*, b^*)$  distribution with

$$a^* = \sum_{i=1}^n X_i + a; \text{ and } b^* = \frac{1}{n + \frac{1}{b}}$$

Hence show that the Bayes estimator,  $\hat{\theta}_B = E(\theta|X)$ , is a weighted sum of the sample mean,  $\bar{X}$  and the prior distribution mean,  $\mu_p$ , i.e.,

$$\hat{\theta}_B = w\bar{X} + (1-w)\mu_p$$

with the weight,  $w = nb/(nb+1)$ .

(ii) Now show that:

$$Var(\hat{\theta}_B) < Var(\bar{X})$$

always, but especially when  $n$  is small.

**14.35** The first 10 samples of the number of “inclusions” on 1-sq meter glass sheets, shown below as the set  $I_{10}$ , has been extracted from the full data set in Table 1.2.

$$I_{10} = \{0, 1, 1, 1, 0, 0, 1, 0, 2, 2\}$$

Consider this as a sample from a Poisson population with true population parameter  $\theta = 1$ .

(i) Use the first five entries to obtain a maximum likelihood estimate of  $\theta$ ; compute the sample variance. Then refer to Exercise 14.34 and use the results there to obtain the Bayes estimate, again using the first five entries along with a prior distribution chosen as a Gamma  $\gamma(2, 1)$ . Obtain the variance of this Bayes estimate. Clearly the sample size is too small, but still use the Gaussian approximation to obtain *approximate* 95% confidence intervals around these estimates. Compare these two different estimates to the true value.

(ii) Repeat (i) this time using the entire 10 samples.

## APPLICATION PROBLEMS

**14.36** A cohort of 100 patients under the age of 35 years (the “Younger” group), and another cohort of the same size, but 35 years and older (the “Older” group), participated in a clinical study where each patient received five embryos in an in-vitro fertilization (IVF) treatment cycle. The result from “Assisted Reproductive Technologies” clinic where the study took place is shown in the table below. The data shows  $x$ , the number of live births per delivered pregnancy, along with how many in each group had the pregnancy outcome of  $x$ .

| $x$<br>No. of live<br>births in a<br>delivered<br>pregnancy | $y_O$  |                            | $y_Y$  |                            |
|---|--|----------------------------|--|----------------------------|
|   | Total no. of<br>“older patients”<br>(out of 100) | with pregnancy outcome $x$ | Total no. of<br>“younger patients”<br>(out of 100) | with pregnancy outcome $x$ |
| 0   | 32   |                            | 8  |                            |
| 1   | 41   |                            | 25   |                            |
| 2   | 21   |                            | 35   |                            |
| 3   | 5  |                            | 23   |                            |
| 4   | 1  |                            | 8  |                            |
| 5   | 0  |                            | 1  |                            |

On the postulate that these data represent random samples from the binomial  $Bi(n, \theta_O)$  population for the “Older” group, and  $Bi(n, \theta_Y)$  for the “Younger” group, obtain 95% confidence interval estimates of both parameters,  $\theta_O$  and  $\theta_Y$ .

Physiologically, these parameters represent the single embryo probability of “success” (i.e., resulting in a live birth at the end of the treatment cycle) for the patients in each group. Comment on whether or not the results of this clinical study indicate that these cohort groups have different IVF treatment success rates, on average.

**14.37** The number of contaminant particles (flaws) found on each standard size silicon wafer produced at a certain manufacturing site is a random variable,  $X$ . In order to characterize this random variable, a random sample of 30 silicon wafers selected by a quality control engineer and examined for flaws resulted in the data shown in the table below, a record of the number of flaws found on each wafer.

|   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|
| 4 | 1 | 2 | 3 | 2 | 1 | 2 | 4 | 0 | 1 |
| 3 | 0 | 0 | 2 | 3 | 0 | 3 | 2 | 1 | 2 |
| 3 | 4 | 1 | 1 | 2 | 2 | 5 | 3 | 1 | 1 |

Postulate an appropriate theoretical probability model for this random variable and estimate the unknown model parameter(s). Include an estimate of the precision of the estimated parameters. State any assumptions you may need to make in answering this question.

**14.38** The following data set, from a study by Lucas (1985)<sup>2</sup> (see also Exercises 12.3, 12.25 and 13.22), shows the number of accidents occurring per quarter (three

<sup>2</sup>Lucas J. M., (1985). “Counted Data CUSUMs,” *Technometrics*, 27, 129–144.

months), over a 10-year period, at a DuPont company facility, separated into two periods: Period I is the first five-year period of the study; Period II, the second five-year period.

| Period I |   |    |    | Period II |   |   |   |
|----------|---|----|----|-----------|---|---|---|
| 5        | 5 | 10 | 8  | 3         | 4 | 2 | 0 |
| 4        | 5 | 7  | 3  | 1         | 3 | 2 | 2 |
| 2        | 8 | 6  | 9  | 7         | 7 | 1 | 4 |
| 5        | 6 | 5  | 10 | 1         | 2 | 2 | 1 |
| 6        | 3 | 3  | 10 | 4         | 4 | 4 | 4 |

- (i) Consider that the entire data set constitutes a random sample of size 40 from a single Poisson population with unknown parameter. Estimate the parameter with a 95% confidence interval.
- (ii) Now consider the data for each period as representing two different random samples of size 20 each, from two different Poisson populations, with unknown parameters  $\theta_1$  for Period I,  $\theta_2$  for Period II. Estimate these parameters with separate, approximate 95% confidence intervals. Compare these two interval estimates and comment on whether or not these two populations are indeed different. If the populations appear different, what do you think may have happened between Period I and Period II at the DuPont company facility that was the site of this study?

**14.39** An exotic flu virus with a long incubation period reappears every year during the long flu season. Unfortunately, there is only a probability  $p$  that an infected patient will show symptoms within the first month; as such the early symptomatic patients constitute only the “leading edge” of the infected members of the population. Assuming, for the sake of simplicity, that once infected, the total number of infected individuals does not increase (the virus is minimally contagious), and assuming that all symptomatic patients eventually come to the same hospital, the following data was obtained over a period of five years by an epidemiologist working with the local hospital doctors.  $N_E$  is the number of “early” symptomatic patients;  $N_T$  is the *total* number of infected patients treated that year. The unknown probability  $p$  is to be determined from this data in order to enable doctors to prepare for the virus’s reappearance the following year.

| Year | Early                 | Total             |
|------|-----------------------|-------------------|
|      | Symptomatics<br>$N_E$ | Infected<br>$N_T$ |
| 1    | 5                     | 7                 |
| 2    | 3                     | 8                 |
| 3    | 3                     | 10                |
| 4    | 7                     | 7                 |
| 5    | 2                     | 8                 |

- (i) Why is this a negative binomial phenomenon? Determine the values of the negative binomial parameter  $k$  and the random variable  $X$  from this data set.
- (ii) Obtain an expression for the maximum likelihood estimate of  $p$  in terms of a general random sample of  $k_i, X_i; i = 1, 2, \dots, n$ . Why is it not possible to use the method of moments to estimate  $p$  in this case?
- (iii) Determine from the data an actual estimate  $\hat{p}$ . Use this estimate to generate a  $7 \times 7$  table of probabilities  $f(x|k)$  for values of  $k = 1, 2, 3, 4, 5, 6, 7$  and

$x = 0, 1, 2, 4, 5, 6$ . Convert this table to a table of  $N_E$  values versus probabilities of observing  $N_T$  infected patients every cycle, for each given  $N_E$ .

**14.40** The data below, taken from Greenwood and Yule, (1920)<sup>3</sup>, shows the frequency of accidents occurring, over a five-week period, to 647 women making high explosives during World War I.

| Number of Accidents | Observed Frequency |
|---------------------|--------------------|
| 0                   | 447                |
| 1                   | 132                |
| 2                   | 42                 |
| 3                   | 21                 |
| 4                   | 3                  |
| 5+                  | 2                  |

- (i) If  $X$  is the random variable representing the number of accidents, determine the mean and the variance of this clearly Poisson-like random variable, and confirm that this is an over-dispersed Poisson variable for which the Poisson population parameter, rather than being constant, varies across the population.
- (ii) For over-dispersed Poisson variables such as this, the more appropriate model is the negative binomial  $NBi(\alpha, p)$ . Determine the method of moments estimates for the unknown parameters  $\alpha$  and  $p$  for this data set.

**14.41** The table below shows the time in months between occurrences of safety violations for three operators, “A,” “B,” and “C,” working in a toll manufacturing facility.

|          |      |      |      |      |      |      |      |      |      |      |
|----------|------|------|------|------|------|------|------|------|------|------|
| <i>A</i> | 1.31 | 0.15 | 3.02 | 3.17 | 4.84 | 0.71 | 0.70 | 1.41 | 2.68 | 0.68 |
| <i>B</i> | 1.94 | 3.21 | 2.91 | 1.66 | 1.51 | 0.30 | 0.05 | 1.62 | 6.75 | 1.29 |
| <i>C</i> | 0.79 | 1.22 | 0.65 | 3.90 | 0.18 | 0.57 | 7.26 | 0.43 | 0.96 | 3.76 |

Postulate an appropriate probability model for the phenomenon in question, treat the data set as three random samples of size  $n = 10$  each, from the three different populations represented by each operator. Obtain precise 95% confidence interval estimates of the unknown population parameters. Interpret your results in terms of any differences that might exist between the safety performances of the three operators.

**14.42** The data set in the table below is the time (in months) from receipt to publication (sometimes known as *time-to-publication*) of 85 papers published in the January 2004 issue of a leading chemical engineering research journal.

<sup>3</sup>Greenwood M. and Yule, G. U. (1920) “An enquiry into the nature of frequency distributions representative of multiple happenings with particular reference of multiple attacks of disease or of repeated accidents.” *Journal Royal Statistical Society* 83:255-279.

|      |      |      |      |      |
|------|------|------|------|------|
| 19.2 | 15.1 | 9.6  | 4.2  | 5.4  |
| 9.0  | 5.3  | 12.9 | 4.2  | 15.2 |
| 17.2 | 12.0 | 17.3 | 7.8  | 8.0  |
| 8.2  | 3.0  | 6.0  | 9.5  | 11.7 |
| 4.5  | 18.5 | 24.3 | 3.9  | 17.2 |
| 13.5 | 5.8  | 21.3 | 8.7  | 4.0  |
| 20.7 | 6.8  | 19.3 | 5.9  | 3.8  |
| 7.9  | 14.5 | 2.5  | 5.3  | 7.4  |
| 19.5 | 3.3  | 9.1  | 1.8  | 5.3  |
| 8.8  | 11.1 | 8.1  | 10.1 | 10.6 |
| 18.7 | 16.4 | 9.8  | 10.0 | 15.2 |
| 7.4  | 7.3  | 15.4 | 18.7 | 11.5 |
| 9.7  | 7.4  | 15.7 | 5.6  | 5.9  |
| 13.7 | 7.3  | 8.2  | 3.3  | 20.1 |
| 8.1  | 5.2  | 8.8  | 7.3  | 12.2 |
| 8.4  | 10.2 | 7.2  | 11.3 | 12.0 |
| 10.8 | 3.1  | 12.8 | 2.9  | 8.8  |

Postulate an appropriate probability model for this random variable, justifying your choice clearly but succinctly. Consider this data as a random sample from the population in question and determine estimates of the unknown population parameters using whatever technique is most convenient. Plot on the same graph, a theoretical model prediction against the data histogram and comment on your model fit to the data. In particular compare the model prediction of the “most popular” time-to-publication,  $x^*$  with the corresponding data value; also compare the probability that a paper will take longer than  $x^*$  months to publish with the proportion of papers from the data table that took longer than  $x^*$  months to publish.

**14.43** The data table below shows  $x$ , a count of the number of species,  $x = 1, 2, \dots, 24$ , and the associated number of Malayan butterflies that have  $x$  number of species. In Fisher *et al.*, (1943)<sup>4</sup>, where the data was first published and analyzed, it was proposed that the appropriate model for the phenomenon is the *logarithmic series distribution* (see Exercise 8.13), with the pdf:

$$f(x) = \frac{\alpha p^x}{x}; 0 < p < 1; x = 1, 2, \dots,$$

where

$$\alpha = \frac{-1}{\ln(1-p)}$$

---

<sup>4</sup>Fisher, R. A., S. Corbet, and C. B. Williams. (1943). “The relation between the number of species and the number of individuals in a random sample of an animal population.” *Journal of Animal Ecology*, 1943: 4258.

|                    |     |    |    |    |    |    |    |    |
|--------------------|-----|----|----|----|----|----|----|----|
| $x$                | 1   | 2  | 3  | 4  | 5  | 6  | 7  | 8  |
| No of species      |     |    |    |    |    |    |    |    |
| $\Phi(x)$          | 118 | 74 | 44 | 24 | 29 | 22 | 20 | 19 |
| Observed Frequency |     |    |    |    |    |    |    |    |
| $x$                | 9   | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| No of species      |     |    |    |    |    |    |    |    |
| $\Phi(x)$          | 20  | 15 | 12 | 14 | 6  | 12 | 6  | 9  |
| Observed Frequency |     |    |    |    |    |    |    |    |
| $x$                | 17  | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| No of species      |     |    |    |    |    |    |    |    |
| $\Phi(x)$          | 9   | 6  | 10 | 10 | 11 | 5  | 3  | 3  |
| Observed Frequency |     |    |    |    |    |    |    |    |

It can be shown that for this random variable and its pdf,

- $E(X) = \alpha p / (1 - p)$
- $Var(X) = \alpha p(1 - \alpha p)(1 - p)^{-2}$

Show that the MLE of the unknown population parameter  $p$  and the method of moments estimator based on the first moment coincide. Obtain an estimate of this population parameter for this data set. Compare the model prediction to the data.

**14.44** The data in the table below, on the wall thickness (in ins) of cast aluminum cylinder heads used in aircraft engine cooling jackets, is from Mee (1990)<sup>5</sup>.

|       |       |       |       |       |       |       |       |       |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 0.223 | 0.228 | 0.214 | 0.193 | 0.223 | 0.213 | 0.218 | 0.215 | 0.233 |
| 0.201 | 0.223 | 0.224 | 0.231 | 0.237 | 0.217 | 0.204 | 0.226 | 0.219 |

(i) Consider this as a random sample from a normal population with unknown parameters  $\mu$  and  $\sigma^2$ . Determine 95% confidence interval estimates of both the mean and the variance of the wall thickness.

(ii) If a customer requires that these wall thicknesses be made to within the specifications,  $0.220 \pm 0.030$  ins, what is the probability that the manufacturer will meet these specifications? State any assumptions required in answering this question.

**14.45** The intrinsic variations in the measured amount of pollution contained in water samples from rivers and streams in a mining region of West Central United States is known to be a normally distributed random variable with a fairly stable standard deviation of 5 milligrams of solids per liter. As an EPA inspector who wishes to test a random selection of  $n$  water samples in order to determine the mean daily rate of pollution within  $\pm 1$  milligram per liter with 95% confidence, how many water samples will you need to take? A selection of 6 randomly selected water samples returned a mean value of 40 milligrams per liter and what seems like an excessive variance of 30 (mg/liter)<sup>2</sup>. Determine a 95% confidence interval around this estimate of  $\sigma^2$  and comment on whether or not this value is excessive compared to the assumed population value.

**14.46** The time (in weeks) between occurrences of “minor safety incidents” in a

---

<sup>5</sup>Mee, R. W., (1990). “An improved procedure for screening based on a correlated, normally distributed variable,” *Technometrics*, 32, 331–337.

Research and Development laboratory is known to be a random variable  $X$  with an exponential distribution

$$f(x) = \frac{1}{\beta} e^{-x/\beta}; 0 < x < \infty$$

where the characteristic parameter,  $\beta$ , is to be determined from a random sample  $X_1, X_2, \dots, X_n$ , using Bayesian principles.

(i) Replace  $1/\beta$  with  $\theta$  in the pdf, and use the following gamma  $\gamma(a, b)$  distribution as the prior distribution for  $\theta$ :

$$f(\theta) = C\theta^{a-1}e^{-\theta/b}$$

where  $C$  is the usual normalization constant, and  $a, b$  are known constants. Obtain the posterior distribution  $f(\theta|x_1, x_2, \dots, x_n)$  and from this, obtain an expression for  $\hat{\theta}_B$ , the posterior mean (i.e., the Bayes' estimate of  $\theta$ ); from this obtain  $\hat{\beta}_B$ , the corresponding estimate of  $\beta$ . Also obtain  $\theta^*$ , the MAP estimate, and from this obtain  $\hat{\beta}^*$  the corresponding estimate of  $\beta$ . Compare these two estimates of  $\beta$ .

(ii) The data set shown here was extracted from recent safety incidents records of the laboratory discussed above. It shows the time (in weeks) between occurrences of the last 11 "minor safety incidents."

|      |      |      |      |      |
|------|------|------|------|------|
| 1.31 | 0.15 | 3.02 | 3.17 | 4.84 |
| 0.71 | 0.70 | 1.41 | 2.68 | 0.68 |

Considering this data set as a specific realization of the random sample  $X_1, X_2, \dots, X_{10}$ , obtain the maximum likelihood estimate,  $\hat{\beta}_{ML}$  of the characteristic parameter  $\beta$ . Employing as a prior distribution for  $\theta = 1/\beta$ , the Gamma distribution given in (i) above, with parameters  $a = 2, b = 1$ , obtain the Bayes' estimate  $\hat{\beta}_B$ , as well as the MAP estimate,  $\hat{\beta}^*$ . Compare these three estimates.

(iii) Plot on one graph, a gamma  $\gamma(2, 1)$  and a gamma  $\gamma(3, 1)$  pdf; compare them. Now repeat part (ii) above but this time use a prior distribution with parameters  $a = 3, b = 1$ . Compare the Bayes and MAP estimates obtained here with those obtained in (ii) and comment on the effect of the prior distributions on these estimates.

**14.47** The number of contaminant particles (flaws) found on each standard size silicon wafer produced at a certain manufacturing site is a Poisson distributed random variable,  $X$ , with an unknown mean  $\theta$ . To estimate this parameter, a sample of  $n$  wafers from a manufactured lot was examined and the number of contaminant particles found on each wafer recorded.

(i) Given that this data set constitutes a random sample  $X_1, X_2, \dots, X_n$ , Use a gamma  $\gamma(\alpha, \beta)$  distribution as a prior pdf for the unknown parameter,  $\theta$ , and obtain  $\hat{\theta}_B$ , the Bayes' estimate for  $\theta$  (i.e. the posterior mean) and show that if  $\hat{\theta}_{ML}$  is the *maximum likelihood* estimate for  $\theta$ , then,

$$\lim_{\substack{\alpha \rightarrow 0 \\ \beta \rightarrow \infty}} \hat{\theta}_B = \hat{\theta}_{ML}$$

(ii) A sample of 30 silicon wafers was examined for flaws and the result (the number of flaws found on each wafer) is displayed in the table below.

|   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|
| 3 | 1 | 2 | 3 | 2 | 1 | 2 | 2 | 0 | 1 |
| 4 | 0 | 0 | 2 | 0 | 0 | 0 | 2 | 1 | 2 |
| 3 | 4 | 1 | 1 | 2 | 2 | 5 | 3 | 1 | 1 |

Use a gamma  $\gamma(2, 2)$  distribution as the prior distribution for  $\theta$  and obtain, using this data set:

- (a) the maximum likelihood estimate;
- (b) the Bayes' estimate, and
- (c) Repeat (a) and (b) using only *the first 10 data points* in the first row of the data table. Comment on how an increase in the number of available data points affects parameter estimation in this particular case.

**14.48** Consider the case in which  $\eta_k$ , the true value of a process variable at time instant  $k$ , is measured as  $y_k$ , i.e.,

$$y_k = \eta_k + \epsilon_k \quad (14.168)$$

where  $\epsilon_k$ , the measurement noise, is usually considered random. The standard procedure for obtaining a good estimate of  $\eta_k$  involves taking repeated measurements and averaging.

However, many circumstances arise in practice when such a strategy is infeasible primarily because of significant process dynamics. Under these circumstances, the process variable changes significantly with time during the period of repeated sampling; the repeated measurements thus provide information about the process variable at *different time instants* and not true replicates of the desired measurement at a specific time instant,  $k$ . Decisions must therefore be made on the true value,  $\eta_k$ , from the *single*, available measurement  $y_k$  — a non-standard problem which may be solved using the Bayesian approach as follows:

(i) **Theory:** Consider  $y_k$  as a realization of the random variable,  $Y_k$ , possessing a normal  $N(\eta_k, \sigma^2)$  distribution with unknown mean  $\eta_k$ , and variance  $\sigma^2$ ; then consider that the (unknown) process dynamics can be approximated by the simple “random walk” model:

$$\eta_k = \eta_{k-1} + w_k \quad (14.169)$$

where  $w_k$ , the “process noise,” is a sequence of independent realizations of the zero mean, Gaussian random variable,  $W$ , with a variance  $v^2$ ; i.e.,  $W \sim N(0, v^2)$ .

This process dynamic model is equivalent to declaring that  $\eta_k$ , the unknown true mean value of  $Y_k$ , has a prior pdf  $N(\eta_{k-1}, v^2)$ ; the measurement equation above, Eq (14.168), implies that the sampling distribution for  $Y_k$  is given by:

$$f(y_k | \eta_k) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left\{ \frac{-(y_k - \eta_k)^2}{2\sigma^2} \right\}$$

Combine this with the prior distribution and obtain an expression for the posterior distribution  $f(\eta_k | y_k)$  and show that the result is a Gaussian pdf with mean  $\tilde{\eta}_k$ , variance  $\tilde{\sigma}^2$  given by:

$$\tilde{\eta}_k = \alpha \eta_{k-1} + (1 - \alpha) y_k \quad (14.170)$$

$$\alpha = \frac{\sigma^2}{\sigma^2 + v^2} \quad (14.171)$$

and

$$\tilde{\sigma}^2 = \frac{1}{\frac{1}{\sigma^2} + \frac{1}{v^2}} \quad (14.172)$$

Thus, by adopting the penultimate Bayes' estimate  $\tilde{\eta}_{k-1}$  as an estimate for  $\eta_{k-1}$  (since it is also unknown) obtain the recursive formula:

$$\tilde{\eta}_k = \alpha \tilde{\eta}_{k-1} + (1 - \alpha) y_k \quad (14.173)$$

for estimating the true value  $\eta_k$  from the single data point  $y_k$ . *This expression is recognizable to engineers as the popular discrete, (first order) exponential filter, for which  $\alpha$  is usually taken as a “tuning parameter”.*

(ii) **Application:** Apply the result in part (i) above to “filter” the following raw data, representing 25 hourly measurements of a polymer product’s solution viscosity (in scaled, coded units), using  $\alpha = 0.20$ , and the initial condition  $\eta_1^* = 20.00$ .

| $k$ | $y_k$ | $k$ | $y_k$ |
|-----|-------|-----|-------|
| 1   | 20.82 | 14  | 18.65 |
| 2   | 20.92 | 15  | 21.48 |
| 3   | 21.46 | 16  | 21.85 |
| 4   | 22.15 | 17  | 22.34 |
| 5   | 19.76 | 18  | 20.35 |
| 6   | 21.91 | 19  | 20.32 |
| 7   | 22.13 | 20  | 22.10 |
| 8   | 24.26 | 21  | 20.69 |
| 9   | 20.26 | 22  | 19.74 |
| 10  | 20.35 | 23  | 20.27 |
| 11  | 18.32 | 24  | 23.33 |
| 12  | 19.24 | 25  | 19.69 |
| 13  | 19.99 |     |       |

Compare a time sequence plot of the resulting filtered value,  $\eta_k^*$ , with that of the raw data. Repeat with  $\alpha = 0.80$  (and the same initial condition) and comment on which filter parameter value (0.20 or 0.80) provides estimates that are “more representative” of the dynamic behavior exhibited by the raw data.

**14.49** Padgett and Spurrier (1990)<sup>6</sup> obtained the following data set for the breaking strengths (in GPa) of carbon fibers used in making composite materials.

|     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 1.4 | 3.7 | 3.0 | 1.4 | 1.0 | 2.8 | 4.9 | 3.7 | 1.8 | 1.6 |
| 3.2 | 1.6 | 0.8 | 5.6 | 1.7 | 1.6 | 2.0 | 1.2 | 1.1 | 1.7 |
| 2.2 | 1.2 | 5.1 | 2.5 | 1.2 | 3.5 | 2.2 | 1.7 | 1.3 | 4.4 |
| 1.8 | 0.4 | 3.7 | 2.5 | 0.9 | 1.6 | 2.8 | 4.7 | 2.0 | 1.8 |
| 1.6 | 1.1 | 2.0 | 1.6 | 2.1 | 1.9 | 2.9 | 2.8 | 2.1 | 3.7 |

It is known that this phenomenon is well-modeled by the Weibull  $W(\zeta, \beta)$  distribution; the objective is to determine values for these unknown population parameters from this sample data, considered as a random sample with  $n = 50$ . However,

<sup>6</sup>Padgett, W.J. and J. D. Spurrier, (1990). Shewhart-type charts for percentiles of strength distributions. *J of Quality Tech.* 22, 283–388.

obtaining Weibull population parameters from sample data is particularly difficult. The following is a method based on the cumulative distribution function, cdf,  $F(x)$ .

From the Weibull pdf, (and from the derivation presented in Chapter 9), first show that the Weibull cdf is given by

$$F(x) = 1 - e^{(-x/\beta)^\zeta} \quad (14.174)$$

and hence show that

$$\zeta(\ln x - \ln \beta) = \ln \{\ln[1 - F(x)]\} \quad (14.175)$$

Observe therefore that given  $F(x)$  and  $x$ , a plot of  $\ln \{\ln[1 - F(x)]\}$  versus  $\ln x$  should result in a straight line with slope  $= \zeta$  and intercept  $= \zeta \ln \beta$ , from which the appropriate values may be determined for the two unknown parameters.

Employ the outlined technique to determine from the supplied data your best estimate of the unknown parameters. Compare your results to the “true” values  $\beta = 2.5$  and  $\zeta = 2.0$  used by Padgett and Spurrier in their analysis.

**14.50** Polygraphs, the so-called “lie-detector” machines based on physiological measurements such as blood pressure, respiration and perspiration, are used frequently in government agencies and other businesses where employees handle highly classified information. While polygraph test results are sometimes permitted in some state courts, they are *not* admissible in federal courts in part because of potential errors and the implications of such errors on the fairness of the justice system.

Since the basic premise of these machines is the measurement of human physiological variables, it is possible to evaluate the performance of polygraphs in somewhat the same manner as one would any other medical diagnostic machine. (See Phillips, Brett, and Beary, (1986)<sup>7</sup> for one such study carried out by a group of physicians.)

The data shown below is a compilation of the result of an extensive study (similar to the Phillips *et al.*, study) in which a group of volunteers were divided into two equal-numbered subgroups of “truth tellers” and “liars.” The tests were repeated 56 times over a period of two weeks and the results tabulated as shown:  $X_A$  is the fraction of the “truth tellers” falsely identified as “liars” by a Type A polygraph machine (i.e., false positives);  $X_B$  is the set of corresponding results for the same subjects, under conditions as close to identical as possible using a Type B polygraph machine. Conversely,  $Y_A$  is the fraction of “liars” misidentified as “truth-tellers” (i.e., false negatives) by the Type A machine with  $Y_B$  as the corresponding results using the Type B machine.

Postulate a reasonable probability model for the random phenomenon in question, providing a brief but adequate justification for your choice. Estimate the model parameters for the four populations and discuss how well your model fits the data.

---

<sup>7</sup>M. Phillips, A. Brett, and J. Beary, (1986). “Lie Detectors Can Make a Liar Out of You,” *Discover*, June 1986, p. 7

| Polygraph Data |       |       |       |
|----------------|-------|-------|-------|
| $X_A$          | $Y_A$ | $X_B$ | $Y_B$ |
| 0.128          | 0.161 | 0.161 | 0.064 |
| 0.264          | 0.117 | 0.286 | 0.036 |
| 0.422          | 0.067 | 0.269 | 0.214 |
| 0.374          | 0.158 | 0.380 | 0.361 |
| 0.240          | 0.105 | 0.498 | 0.243 |
| 0.223          | 0.036 | 0.328 | 0.235 |
| 0.281          | 0.210 | 0.159 | 0.024 |
| 0.316          | 0.378 | 0.391 | 0.114 |
| 0.341          | 0.283 | 0.154 | 0.067 |
| 0.397          | 0.166 | 0.216 | 0.265 |
| 0.037          | 0.212 | 0.479 | 0.378 |
| 0.097          | 0.318 | 0.049 | 0.004 |
| 0.112          | 0.144 | 0.377 | 0.043 |
| 0.216          | 0.281 | 0.327 | 0.271 |
| 0.265          | 0.238 | 0.563 | 0.173 |
| 0.225          | 0.043 | 0.169 | 0.040 |
| 0.253          | 0.200 | 0.541 | 0.410 |
| 0.211          | 0.299 | 0.338 | 0.031 |
| 0.301          | 0.106 | 0.438 | 0.131 |
| 0.469          | 0.161 | 0.242 | 0.023 |
| 0.410          | 0.151 | 0.461 | 0.159 |
| 0.454          | 0.200 | 0.694 | 0.265 |
| 0.278          | 0.129 | 0.439 | 0.013 |
| 0.236          | 0.222 | 0.194 | 0.190 |
| 0.118          | 0.245 | 0.379 | 0.030 |
| 0.109          | 0.308 | 0.368 | 0.069 |
| 0.035          | 0.019 | 0.426 | 0.127 |
| 0.269          | 0.146 | 0.597 | 0.144 |

| Polygraph Data |       |       |       |
|----------------|-------|-------|-------|
| $X_A$          | $Y_A$ | $X_B$ | $Y_B$ |
| 0.175          | 0.368 | 0.441 | 0.024 |
| 0.425          | 0.327 | 0.412 | 0.218 |
| 0.119          | 0.698 | 0.295 | 0.057 |
| 0.380          | 0.054 | 0.136 | 0.081 |
| 0.234          | 0.070 | 0.438 | 0.085 |
| 0.323          | 0.057 | 0.445 | 0.197 |
| 0.356          | 0.506 | 0.239 | 0.111 |
| 0.401          | 0.142 | 0.207 | 0.011 |
| 0.444          | 0.356 | 0.251 | 0.029 |
| 0.326          | 0.128 | 0.430 | 0.229 |
| 0.484          | 0.108 | 0.195 | 0.546 |
| 0.280          | 0.281 | 0.429 | 0.039 |
| 0.435          | 0.211 | 0.581 | 0.061 |
| 0.172          | 0.333 | 0.278 | 0.136 |
| 0.235          | 0.100 | 0.151 | 0.014 |
| 0.418          | 0.114 | 0.374 | 0.055 |
| 0.366          | 0.083 | 0.638 | 0.031 |
| 0.077          | 0.251 | 0.187 | 0.239 |
| 0.352          | 0.085 | 0.680 | 0.106 |
| 0.231          | 0.225 | 0.198 | 0.066 |
| 0.175          | 0.325 | 0.533 | 0.132 |
| 0.290          | 0.352 | 0.187 | 0.240 |
| 0.099          | 0.185 | 0.340 | 0.070 |
| 0.254          | 0.287 | 0.391 | 0.197 |
| 0.556          | 0.185 | 0.318 | 0.071 |
| 0.407          | 0.109 | 0.102 | 0.351 |
| 0.191          | 0.049 | 0.512 | 0.072 |
| 0.232          | 0.076 | 0.356 | 0.048 |

TABLE 14.1: Summary of estimation results

| Population Parameter        | Point Estimator                                    | Expected Value    | Variance            | Conf. Interval Estimator  |
|-----------------------------|--|-------------------|---------------------|---|
| $\theta$                    | $\hat{\Theta}$                                     | $E(\hat{\Theta})$ | $Var(\hat{\Theta})$ | $(1 - \alpha) \times 100\%$                                       |
| Mean, $\mu$<br>( $n < 30$ ) | $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$             | $\mu$             | $\sigma^2/n$        | $\bar{X} \pm z_{\alpha/2} \left( \frac{\sigma}{\sqrt{n}} \right)$ |
| Variance, $\sigma^2$        | $S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$ | $\sigma^2$        | $s^2/n$             | $\bar{X} \pm t_{\alpha/2}(n-1) \left( \frac{s^2}{n} \right)$      |
| Binomial Proportion, $p$    | $\frac{\bar{X}}{n}$                                | $p$               | $pq/n$              | $\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$        |

**TABLE 14.2:** Some population parameters and conjugate prior distributions appropriate for their Bayesian estimation

| Population and Parameter          | Sampling Distribution<br>$f(x \theta)$   | Conjugate prior Distribution<br>$f(\theta)$   |
|-----------------------------------|--|---|
| Binomial<br>$\theta = p$          | $C_1 \theta^x (1-\theta)^{n-x}$  | Beta, $B(\alpha, \beta)$<br>$C_2 \theta^{\alpha-1} (1-\theta)^{\beta-1}$                                    |
| Poisson<br>$\theta = \lambda$     | $\frac{\theta^{(\sum_{i=1}^n x_i)} e^{-\theta}}{\prod_{i=1}^n x_i!}$                                     | Gamma, $\gamma(\alpha, \beta)$<br>$C_2 \theta^{\alpha-1} e^{-\theta/\beta}$                                 |
| Exponential<br>$\theta = 1/\beta$ | $\theta e^{-\theta(\sum_{i=1}^n x_i)}$   | Gamma, $\gamma(\alpha, \beta)$<br>$C_2 \theta^{\alpha-1} e^{-\theta/\beta}$                                 |
| Gaussian<br>$\theta = \mu$        | $\frac{1}{\sigma \sqrt{2\pi}} \exp \left\{ \frac{-\sum_{i=1}^n (x_i - \theta)^2}{2\sigma^2} \right\}$    | Gaussian, $N(\tau, v^2)$<br>$\frac{1}{v \sqrt{2\pi}} \exp \left\{ \frac{-(\theta - \tau)^2}{2v^2} \right\}$ |
| Gaussian<br>$\theta = \sigma^2$   | $\frac{1}{(\theta)^{1/2} \sqrt{2\pi}} \exp \left\{ \frac{-\sum_{i=1}^n (x_i - \mu)^2}{2\theta} \right\}$ | Inverse Gamma,<br>$IG(\alpha, \beta)$<br>$\frac{C}{\theta^{\alpha+1}} e^{-\frac{\beta}{\theta}}$            |

# Chapter 15

## Hypothesis Testing

|  |  |     |
|--|--|-----|
| 15.1   | Introduction .....   | 552 |
| 15.2   | Basic Concepts .....   | 553 |
| 15.2.1   | Terminology and Definitions .....                            | 554 |
| Statistical Hypothesis .....                                 | 554  |     |
| Test Statistic, Critical Region and Significance Level ..... | 556  |     |
| Potential Errors, Risks, and Power .....                     | 557  |     |
| Sensitivity and Specificity .....                            | 558  |     |
| The $p$ -value .....   | 559  |     |
| 15.2.2   | General Procedure .....                                      | 560 |
| 15.3   | Concerning Single Mean of a Normal Population .....          | 561 |
| 15.3.1   | $\sigma$ Known; the “z-test” .....                           | 563 |
| Using MINITAB .....  | 567  |     |
| 15.3.2   | $\sigma$ Unknown; the “t-test” .....                         | 570 |
| Using MINITAB .....  | 573  |     |
| 15.3.3   | Confidence Intervals and Hypothesis Tests .....              | 573 |
| 15.4   | Concerning Two Normal Population Means .....                 | 576 |
| 15.4.1   | Population Standard Deviations Known .....                   | 576 |
| 15.4.2   | Population Standard Deviations Unknown .....                 | 578 |
| Equal standard deviations .....                              | 578  |     |
| Using MINITAB .....  | 580  |     |
| Unequal standard deviations .....                            | 581  |     |
| Confidence Intervals and Two-Sample tests .....              | 581  |     |
| An Illustrative Example: The Yield Improvement Problem ..... | 582  |     |
| 15.4.3   | Paired Differences .....                                     | 585 |
| 15.5   | Determining $\beta$ , Power, and Sample Size .....           | 589 |
| 15.5.1   | $\beta$ and Power .....                                      | 591 |
| 15.5.2   | Sample Size .....  | 593 |
| Practical Considerations .....                               | 596  |     |
| 15.5.3   | $\beta$ and Power for Lower-Tailed and Two-Sided Tests ..... | 598 |
| 15.5.4   | General Power and Sample Size Considerations .....           | 599 |
| 15.6   | Concerning Variances of Normal Populations .....             | 600 |
| 15.6.1   | Single Variance .....  | 601 |
| 15.6.2   | Two Variances .....  | 603 |
| 15.7   | Concerning Proportions .....                                 | 606 |
| 15.7.1   | Single Population Proportion .....                           | 607 |
| Large Sample Approximations .....                            | 608  |     |
| Exact Tests .....  | 609  |     |
| 15.7.2   | Two Population Proportions .....                             | 610 |
| 15.8   | Concerning Non-Gaussian Populations .....                    | 612 |
| 15.8.1   | Large Sample Test for Means .....                            | 613 |
| 15.8.2   | Small Sample Tests .....                                     | 614 |
| 15.9   | Likelihood Ratio Tests .....                                 | 616 |
| 15.9.1   | General Principles .....                                     | 616 |
| 15.9.2   | Special Cases .....  | 618 |
|  | Normal Population; Known Variance .....                      | 619 |

|  |     |
|--|-----|
| Normal Population; Unknown Variance .....          | 620 |
| 15.9.3 Asymptotic Distribution for $\Lambda$ ..... | 622 |
| 15.10 Discussion .....                             | 623 |
| 15.11 Summary and Conclusions .....                | 624 |
| REVIEW QUESTIONS .....                             | 626 |
| EXERCISES .....                                    | 629 |
| APPLICATION PROBLEMS .....                         | 637 |

*The great tragedy of science —  
the slaying of a beautiful hypothesis by an ugly fact.*

T. H. Huxley (1825–1895)

Since turning our attention fully to Statistics in Part IV, our focus has been on characterizing the population completely, using finite-sized samples. The discussion that began with sampling in Chapter 13, providing the mathematical foundation for characterizing the variability in random samples, and which continued with estimation in Chapter 14, providing techniques for determining values for populations parameters, concludes in this chapter with hypothesis testing. This final tier of the statistical inference edifice is concerned with making — and testing — assertive statements about the population. Such statements are often necessary to solve practical problems, or to answer questions of practical importance; and this chapter is devoted to presenting the principles, practice and mechanics of testing the validity of hypothesized statements regarding the distribution of populations. The chapter covers extensive ground — from traditional techniques applied to traditional Gaussian problems, to non-Gaussian problems and some non-traditional techniques; it ends with a brief but frank discussion of persistent criticisms of hypothesis tests and some practical recommendations for handling such criticisms.

## 15.1 Introduction

We begin our discussion by returning to the first problem presented in Chapter 1 concerning yields from two chemical processes; we wish to use it to illustrate the central issues with hypothesis testing. Recall that the problem requires that we decide which process, the challenger, A, or the incumbent, B, should be chosen for commercial operation. The decision is to be based on economically driven comparisons that translate to answering the following mathematical questions about the yields  $Y_A$  and  $Y_B$ :

1. Is  $Y_A \geq 74.5$  and  $Y_B \geq 74.5$ , consistently?
2. Is  $Y_A > Y_B$ ?
3. If yes, is  $Y_A - Y_B > 2$ ?

To deal with the problem systematically, inherent random variability compels us to start by characterizing the populations fully with pdfs which are then used to answer these questions. This requires that we postulate an appropriate probability model, and determine values for the unknown parameters from sample data.

Here is what we know thus far (from Chapters 1 and 12, and from the various examples in Chapter 14): we have plotted histograms of the data and postulated that these are samples from Gaussian-distributed populations; we have computed sample averages,  $\bar{y}_A, \bar{y}_B$ , and sample standard deviations,  $s_A, s_B$ ; and in the various Chapter 14 examples, we have obtained point and interval estimates for the population means  $\mu_A, \mu_B$ , and the population standard deviations  $\sigma_A, \sigma_B$ .

But by themselves, these results are not quite sufficient to answer the questions posed above. To answer the questions, consider the following statements and the implications of being able to confirm or refute them:

1.  $Y_A$  is a random variable characterized by a normal population with mean value 75.5 and standard deviation 1.5, i.e.,  $Y_A \sim N(75.5, 1.5^2)$ ; similarly,  $Y_B \sim N(72.5, 2.5^2)$ ; as a consequence,
2. The random variables,  $Y_A$  and  $Y_B$ , are *not* from the same distribution because  $\mu_A \neq \mu_B$  and  $\sigma_A^2 \neq \sigma_B^2$ ; in particular,  $\mu_A > \mu_B$ ;
3. Furthermore,  $\mu_A - \mu_B > 2$ .

This is a collection of assertions about these two populations, statements which, if confirmed, will enable us answer the questions raised. For example, Statement #1 will allow us to answer Question 1 by making it possible to compute the probabilities  $P(Y_A \geq 74.5)$  and  $P(Y_B \geq 74.5)$ ; Statement #2 will allow us to answer Question 2, and Statement #3, Question 3. How practical problems are formulated as statements of this type, and how such statements are confirmed or refuted, all fall under the formal subject matter of hypothesis testing. In general, the validity of such statements (or other assumptions about the population from which the sample data were obtained) is checked by

1. Proposing an appropriate “statistical hypothesis” about the problem at hand; and
2. Testing this hypothesis against the evidence contained in the data.

## 15.2 Basic Concepts

### 15.2.1 Terminology and Definitions

Before launching into a discussion of the principles and mechanics of hypothesis testing, it is important to introduce first some terminology and definitions.

#### Statistical Hypothesis

A statistical hypothesis is a statement (an assertion or postulate) about the distribution of one or more populations. (Theoretically, the statistical hypothesis is a statement regarding one or more postulated distributions for the random variable  $X$  — distributions for which the statement is presumed to be true. A *simple* hypothesis specifies a single distribution for  $X$ ; a *composite* hypothesis specifies more than one distribution for  $X$ .)

Modern hypothesis testing involves two hypotheses:

1. The *null hypothesis*,  $H_0$ , which represents the primary, “status quo” hypothesis that we are predisposed to believe as true (a plausible explanation of the observation) unless there is evidence in the data to indicate otherwise — in which case, it will be rejected in favor of a postulated alternative.
2. The *alternative hypothesis*,  $H_a$ , the carefully defined complement to  $H_0$  that we are willing to consider in replacement if  $H_0$  is rejected.

For example, the portion of Statement #1 above concerning  $Y_A$  may be formulated more formally as:

$$\begin{aligned} H_0 : \mu_A &= 75.5 \\ H_a : \mu_A &\neq 75.5 \end{aligned} \tag{15.1}$$

The implication here is that we are willing to entertain the fact that the true value of  $\mu_A$ , the mean value of the yield obtainable from process A, is 75.5; that any deviation of the sample data average from this value is due to purely random variability and is not significant (i.e., that this postulate explains the observed data). The alternative is that any observed difference between the sample average and 75.5 is *real* and not just due to random variability; that the alternative provides a better explanation of the data. Observe that this alternative makes no distinction between values that are less than 75.5 or greater; so long as there is evidence that the observed sample average is different from 75.5 (whether greater than or less than),  $H_0$  is to be rejected in favor of this  $H_a$ . Under these circumstances, since the alternative admits of values of  $\mu_A$  that can be less than 75.5 or greater than 75.5, it is called a *two-sided* hypothesis.

It is also possible to formulate the problem such that the alternative actually “chooses sides,” for example:

$$\begin{aligned} H_0 : \mu_A &= 75.5 \\ H_a : \mu_A &< 75.5 \end{aligned} \quad (15.2)$$

In this case, when the evidence in the data does not support  $H_0$  the only other option is that  $\mu_A < 75.5$ . Similarly, if the hypotheses are formulated instead as:

$$\begin{aligned} H_0 : \mu_A &= 75.5 \\ H_a : \mu_A &> 75.5 \end{aligned} \quad (15.3)$$

the alternative, if the equality conjectured by the null hypothesis fails, is that the mean must then be greater. These are *one-sided* hypotheses, for obvious reasons.

A *test* of a statistical hypothesis is a procedure for deciding when to reject  $H_0$ . The conclusion of a hypothesis test is either a decision to *reject*  $H_0$  in favor of  $H_a$  or else to *fail to reject*  $H_0$ . Strictly speaking, one never actually “accepts” a hypothesis; one just fails to reject it.

As one might expect, the conclusion drawn from a hypothesis test is shaped by how  $H_a$  is framed in contrast to  $H_0$ . How to formulate the  $H_a$  appropriately is best illustrated with an example.

#### **Example 15.1: HYPOTHESES FORMULATION FOR COMPARING ENGINEERING TRAINING PROGRAMS**

As part of an industrial training program for chemical engineers in their junior year, some trainees are instructed by Method A, and some by Method B. If random samples of size 10 each are taken from large groups of trainees instructed by each of these two techniques, and each trainee’s score on an appropriate achievement test is shown below, formulate a null hypothesis  $H_0$ , and an appropriate alternative  $H_a$ , to use in testing the claim that Method B is more effective.

|          |    |    |    |    |    |    |    |    |    |    |
|----------|----|----|----|----|----|----|----|----|----|----|
| Method A | 71 | 75 | 65 | 69 | 73 | 66 | 68 | 71 | 74 | 68 |
| Method B | 72 | 77 | 84 | 78 | 69 | 70 | 77 | 73 | 65 | 75 |

#### **Solution:**

We do return to this example later to provide a solution to the problem posed; for now, we address only the issue of formulating the hypotheses to be tested.

Let  $\mu_A$  represent the true mean score for engineers trained by Method A, and  $\mu_B$ , the true mean score for those trained by the other method. The status quo postulate is to presume that there is no difference between the two methods; that any observed difference is due to pure chance alone. The key now is to inquire: if there is evidence in the data that contradicts this status quo postulate, what end result are we interested in testing this evidence against? Since the claim we are

interested in confirming or refuting is that Method B is more effective, then the proper formulation of the hypotheses to be tested is as follows:

$$\begin{aligned} H_0 : \mu_A &= \mu_B \\ H_a : \mu_A &< \mu_B \end{aligned} \quad (15.4)$$

By formulating the problem in this fashion, any evidence that contradicts the null hypothesis will cause us to reject it in favor of something that is actually relevant to the problem at hand.

Note that in this case specifying  $H_a$  as  $\mu_A \neq \mu_B$  does not help us answer the question posed; by the same token, neither does specifying  $H_a$  as  $\mu_A > \mu_B$  because if it is true that  $\mu_A < \mu_B$ , then the evidence in the data will not support the alternative — a circumstance which, by default, will manifest as a misleading lack of evidence to reject  $H_0$ .

Thus, in formulating statistical hypotheses, it is customary to state  $H_0$  as the “no difference,” *nothing-interesting-is-happening* hypothesis; the alternative,  $H_a$ , is then selected to answer the question of interest when there is evidence in the data to contradict the null hypothesis. (See Section 15.10 below for additional discussion about this and other related issues.)

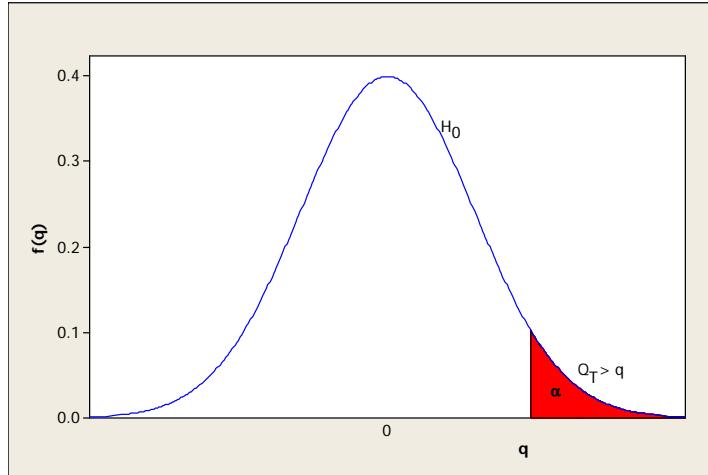
A classic illustration of these principles is the US legal system in which a defendant is considered innocent until proven guilty. In this case, the null hypothesis is that this defendant is no different from any other innocent individual; after evidence has been presented to the jury by the prosecution, the verdict is handed down either that the defendant is guilty (i.e., rejecting the null hypothesis) or the defendant is not guilty (i.e., failing to reject the null hypotheses). Note that the defendant is *not* said to be “innocent;” instead, the defendant is pronounced “not guilty,” which is tantamount to a decision *not* to reject the null hypothesis.

Because hypotheses are statements about populations, and, as with estimation, hypothesis tests are based on finite-sized sample data, such tests are subject to random variability and are therefore only meaningful in a probabilistic sense. This leads us to the next set of definitions and terminology.

### Test Statistic, Critical Region and Significance Level

To test a hypothesis,  $H_0$ , about a population parameter,  $\theta$ , for a random variable,  $X$ , against an alternative,  $H_a$ , a random sample,  $X_1, X_2, \dots, X_n$  is acquired, from which an estimator for  $\theta$ , say  $U(X_1, X_2, \dots, X_n)$ , is then obtained. (Recall that  $U$  is a random variable whose specific value will vary from sample to sample.)

A *test statistic*,  $Q_T(U, \theta)$ , is an appropriate function of the parameter  $\theta$  and its estimator,  $U$ , that will be used to determine whether or not to reject  $H_0$ . (What “appropriate” means will be clarified shortly.) A *critical region* (or rejection region),  $R_C$ , is a region representing the numerical values of the test statistic ( $Q_T > q$ , or  $Q_T < q$ , or both) that will trigger the rejection of  $H_0$ ; i.e., if  $Q_T \in R_C$ ,  $H_0$  will be rejected. Strictly speaking, the critical region is



**FIGURE 15.1:** A distribution for the null hypothesis,  $H_0$ , in terms of the test statistic,  $Q_T$ , where the shaded rejection region,  $Q_T > q$ , indicates a significance level,  $\alpha$

for the random variable,  $X$ ; but since the random sample from  $X$  is usually converted to a test statistic, there is a corresponding mapping of this region by  $Q_T(\cdot)$ ; it is therefore acceptable to refer to the critical region in terms of the test statistic.

Now, because the estimator  $U$  is a random variable, the test statistic will itself also be a random variable, with the following serious implication: *there is a non-zero probability that  $Q_T \in R_C$  even when  $H_0$  is true.* This unavoidable consequence of random variability forces us to design the hypothesis test such that  $H_0$  is rejected only if it is “highly unlikely” for  $Q_T \in R_C$  when  $H_0$  is true. How unlikely is “highly unlikely?” This is quantified by specifying a value  $\alpha$  such that

$$P(Q_T \in R_C | H_0 \text{ true}) \leq \alpha \quad (15.5)$$

with the implication that the probability of rejecting  $H_0$ , when it is in fact true, is never greater than  $\alpha$ . This quantity, often set in advance as a small value (typically, 0.1, 0.05, or 0.01), is called the *significance level* of the test. Thus, the significance level of a test is the upper bound on the probability of rejecting  $H_0$  when it is true; it determines the boundaries of the critical region  $R_C$ .

These concepts are illustrated in Fig 15.1 and lead directly to the consideration of the potential errors to which hypothesis tests are susceptible, the associated risks, and the sensitivity of a test in leading to the correct decision.

### Potential Errors, Risks, and Power

Hypothesis tests are susceptible to two types of errors:

**TABLE 15.1:** Hypothesis test decisions and risks

| <b>DECISION →</b> | <b>Fail to Reject</b>   | <b>Reject</b>  |
|-------------------|---|--|
| <b>TRUTH ↓</b>    | $H_0$   | $H_0$  |
| $H_0$ True        | Correct Decision<br><i>Probability: <math>(1 - \alpha)</math></i> | Type I Error<br><i>Risk: <math>\alpha</math></i>                 |
| $H_a$ True        | Type II Error<br><i>Risk: <math>\beta</math></i>                  | Correct Decision<br><i>Probability: <math>(1 - \beta)</math></i> |

1. *TYPE I error*: the error of rejecting  $H_0$  when it is in fact true. This is the legal equivalent of convicting an innocent defendant.
2. *TYPE II error*: the error of failing to reject  $H_0$  when it is false, the legal equivalent of letting a guilty defendant go scotfree.

Of course a hypothesis test can also result in the correct decision in two ways: rejecting the null hypothesis when it is false, or failing to reject the null hypothesis when it is true.

From the definition of the critical region,  $R_C$ , and the significance level, the probability of committing a Type I error is  $\alpha$ ; i.e.,

$$P(Q_T \in R_C | H_0 \text{ true}) = \alpha \quad (15.6)$$

It is therefore called the  $\alpha$ -risk. The probability of correctly refraining from rejecting  $H_0$  when it is true will be  $(1 - \alpha)$ .

By the same token, it is possible to compute the probability of committing a Type II error. It is customary to refer to this value as  $\beta$ , i.e.,

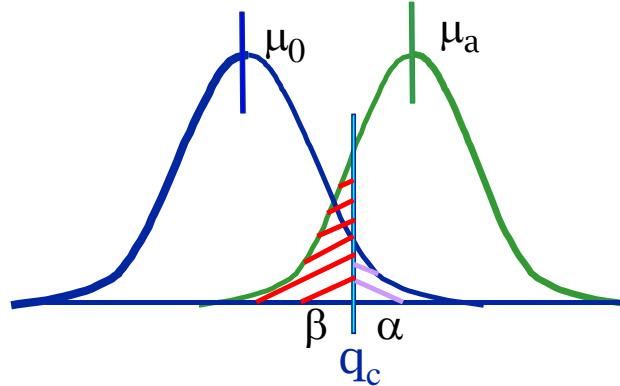
$$P(Q_T \notin R_C | H_0 \text{ false}) = \beta \quad (15.7)$$

so that the probability of committing a Type II error is called the  $\beta$ -risk. The probability of correctly rejecting a null hypothesis that is false is therefore  $(1 - \beta)$ .

It is important now to note that the two correct decisions and the probabilities associated with each one are fundamentally different. Primarily because  $H_0$  is the “status quo” hypothesis, correctly rejecting a null hypothesis,  $H_0$ , that is false is of greater interest because such an outcome indicates that the test has detected the occurrence of something significant. Thus,  $(1 - \beta)$ , the probability of correctly rejecting the false null hypothesis when the alternative hypothesis is true, is known as the *power* of the test. It provides a measure of the sensitivity of the test. These concepts are summarized in Table 15.1 and also in Fig 15.2.

### Sensitivity and Specificity

Because their results are binary decisions (reject  $H_0$  or fail to reject it), hypothesis tests belong in the category of *binary classification tests*; and the



**FIGURE 15.2:** Overlapping distributions for the null hypothesis,  $H_0$  (with mean  $\mu_0$ ), and alternative hypothesis,  $H_a$  (with mean  $\mu_a$ ), showing Type I and Type II error risks  $\alpha$ ,  $\beta$ , along with  $q_c$  the boundary of the critical region of the test statistic,  $Q_T$

effectiveness of such tests are characterized in terms of sensitivity and specificity. The *sensitivity* of a test is the percentage of true “positives” (in this case,  $H_0$  deserving of rejection) that it correctly classifies as such. The *specificity* is the percentage of true “negatives” ( $H_0$  that should *not* be rejected) that is correctly classified as such. Sensitivity therefore measures the ability to identify true positives correctly; specificity, the ability to identify true negatives correctly.

These performance measures are related to the risks and errors discussed previously. If the percentages are expressed as probabilities, then sensitivity is  $(1 - \beta)$ , and specificity,  $(1 - \alpha)$ . The fraction of “false positives” ( $H_0$  that should *not* be rejected but is) is  $\alpha$ ; the fraction of “false negatives” ( $H_0$  that should be rejected but is not) is  $\beta$ . As we show later, for a fixed sample size, improving one measure can only be achieved at the expense of the other, i.e., improvements in specificity must be traded off for a commensurate loss of sensitivity, and vice versa.

### The *p*-value

Rather than fix the significance level,  $\alpha$ , ahead of time, suppose it is free to vary. For any given value of  $\alpha$ , let the corresponding critical/rejection region be represented as  $R_C(\alpha)$ . As discussed above,  $H_0$  is rejected whenever the test statistic,  $Q_T$ , is such that  $Q_T \in R_C(\alpha)$ . For example, from Fig 15.1, the region  $R_C(\alpha)$  is the set of all values of  $Q_T$  that exceed the specific value  $q$ . Observe that as  $\alpha$  decreases, the “size” of the set  $R_C(\alpha)$  also decreases, and vice versa. The *smallest* value of  $\alpha$  for which the specific value of the test statistic  $Q_T(x_1, x_2, \dots, x_n)$  (determined from the data set  $x_1, x_2, \dots, x_n$ ) falls in the critical region (i.e.,  $Q_T(x_1, x_2, \dots, x_n) \in R_C(\alpha)$ ) is known as the *p-value* associated with this data set (and the resulting test statistic). Technically,

therefore, the  $p$ -value is the *smallest* significance level at which  $H_0$  will be rejected given the observed data.

This somewhat technical definition of the  $p$ -value is sometimes easier to understand as follows: given specific observations  $x_1, x_2, \dots, x_n$  and the corresponding test statistic  $Q_T(x_1, x_2, \dots, x_n)$  computed from them to yield the specific value  $q$ ; the  $p$ -value associated with the observations and the corresponding test statistic is defined by the following probability statement:

$$p = P[Q_T(x_1, x_2, \dots, x_n; \theta) \geq q | H_0] \quad (15.8)$$

In words, this is the probability of obtaining the specific test statistic value,  $q$ , or something more extreme, if the null hypothesis is true. Note that  $p$ , being a function of a statistic, is itself a statistic — a subtle point that is often easy to miss; the implication is that  $p$  is itself subject to purely random variability.

Knowing the  $p$ -value therefore allows us to carry out hypotheses tests at any significance level, without restriction to pre-specified  $\alpha$  values. In general, a low value of  $p$  indicates that, given the evidence in the data, the null hypothesis,  $H_0$ , is highly unlikely to be true. This follows from Eq (15.8).  $H_0$  is then rejected at the significance level,  $p$ , which is why the  $p$ -value is sometimes referred to as *the observed significance level* — observed from the sample data, as opposed to being fixed, *a-priori*, at some pre-specified value,  $\alpha$ .

Nevertheless, in many applications (especially in scientific publications), there is an enduring traditional preference for employing fixed significance levels (usually  $\alpha = 0.05$ ). In this case, the  $p$ -value is used to make decisions as follows: if  $p < \alpha$ ,  $H_0$  will be rejected at the significance level  $\alpha$ ; if  $p > \alpha$ , we fail to reject  $H_0$  at the same significance level  $\alpha$ .

### 15.2.2 General Procedure

The general procedure for carrying out modern hypotheses tests is as follows:

1. Define  $H_0$ , the hypothesis to be tested, and pair it with the alternative  $H_a$ , formulated appropriately to answer the question at hand;
2. Obtain sample data, and from it, the test statistic relevant to the problem at hand;
3. Make a decision about  $H_0$  as follows: Either
  - (a) Specify the significance level,  $\alpha$ , at which the test is to be performed, and hence determine the critical region (equivalently, the critical value of the test statistic) that will trigger rejection; then
  - (b) Evaluate the specific test statistic value in relation to the critical region and reject, or fail to reject,  $H_0$  accordingly;

or else,

- (a) Compute the  $p$ -value corresponding to the test statistic, and
- (b) Reject, or fail to reject,  $H_0$  accordingly on this basis.

How this general procedure is applied depends on the specific problem at hand: the nature of the random variable, hence the underlying postulated population itself; what is known or unknown about the population; the particular population parameter that is the subject of the test; and the nature of the question to be answered. The remainder of this chapter is devoted to presenting the principles and mechanics of the various hypothesis tests commonly encountered in practice, some of which are so popular that they have acquired recognizable names (for example the  $z$ -test;  $t$ -test;  $\chi^2$ -test;  $F$ -test; etc). By taking time to provide the *principles* along with the mechanics, our objective is to supply the reader with the sort of information that should help to prevent the surprisingly common mistake of misapplying some of these tests. The chapter closes with a brief discussion of some criticisms and potential shortcomings of classical hypothesis testing.

### 15.3 Concerning Single Mean of a Normal Population

Let us return to the illustrative statements made earlier in this chapter regarding the yields from two competing chemical processes. In particular, let us recall the first half of the statement about the yield of process A — that  $Y_A \sim N(75.5, 1.5^2)$ . Suppose that we are first interested in testing the validity of this statement by inquiring whether or not the true mean of the process yield is 75.5. The starting point for this exercise is to state the null hypothesis, which in this case is:

$$\mu_A = 75.5 \quad (15.9)$$

since 75.5 is the specific postulated value for the unknown population mean  $\mu_A$ . Next, we must attach an appropriate alternative hypothesis. The original statement is a categorical one that  $Y_A$  comes from the distribution  $N(75.5, 1.5^2)$ , with the hope of being able to use this statement to distinguish the  $Y_A$  distribution from the  $Y_B$  distribution. (How this latter task is accomplished is discussed later). Thus, the only alternative we are concerned about, should  $H_0$  prove false, is that the true mean is not equal to 75.5; we do not care if the true mean is less than, or greater than the postulated value. In this case, the appropriate  $H_a$  is therefore:

$$\mu_A \neq 75.5 \quad (15.10)$$

Next, we need to gather “evidence” in the form of sample data from process A. Such data, with  $n = 50$ , was presented in Chapter 1 (and employed in the examples of Chapter 14), from which we have obtained a sample average,

$\bar{y}_A = 75.52$ . And now, the question to be answered by the hypothesis test is as follows: is the observed difference between the postulated true population mean,  $\mu_A = 75.5$ , and the sample average computed from sample process data,  $\bar{y}_A = 75.52$ , due purely to random variation or does it indicate a real (and significant) difference between postulate and data? From Chapters 13 and 14, we now know that answering this question requires a sampling distribution that describes the variability intrinsic to samples. In this specific case, we know that for a sample average  $\bar{X}$  obtained from a random sample of size  $n$  from a  $N(\mu, \sigma^2)$  distribution, the statistic

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad (15.11)$$

has the standard normal distribution, provided that  $\sigma$  is known. This immediately suggests, within the context of hypothesis testing, that the following test statistic:

$$Z = \frac{\bar{y}_A - 75.5}{1.5/\sqrt{n}} \quad (15.12)$$

may be used to test the validity of the hypothesis, for any sample average computed from any sample data set of size  $n$ . This is because we can use  $Z$  and its pdf to determine the critical/rejection region. In particular, by specifying a significance level  $\alpha = 0.05$ , the rejection region is determined as the values  $z$  such that:

$$R_C = \{z|z < -z_{0.025}; z > z_{0.025}\} \quad (15.13)$$

(because this is a two-sided test). From the cumulative probability characteristics of the standard normal distribution, we obtain (using computer programs such as MINITAB)  $z_{0.025} = 1.96$  as the value of the standard normal variate for which  $P(Z > z_{0.025}) = 0.025$ , i.e.,

$$R_C = \{z|z < -1.96; z > 1.96\}; \text{ or } |z| > 1.96 \quad (15.14)$$

The implication: if the specific value computed for  $Z$  from any sample data set exceeds 1.96 in absolute value,  $H_0$  will be rejected.

In the specific case of  $\bar{y}_A = 75.52$  and  $n = 50$ , we obtain a specific value for this test statistic as  $z = 0.094$ . And now, because this value  $z = 0.094$  does not lie in the critical/rejection region defined in Eq (15.14), we conclude that there is no evidence to reject  $H_0$  in favor of the alternative. The data does not contradict the hypothesis.

Alternatively, we could compute the  $p$ -value associated with this test statistic (for example, using the cumulative probability feature of MINITAB):

$$P(z > 0.094 \text{ or } z < -0.094) = P(|z| > 0.094) = 0.925 \quad (15.15)$$

implying that if  $H_0$  is true, the probability of observing, by pure chance alone, the sample average data actually observed, or something “more extreme,” is

very high at 0.925. Thus, there is no evidence in this data set to justify rejecting  $H_0$ . From a different perspective, note that this  $p$ -value is nowhere close to being *lower* than the prescribed significance level,  $\alpha = 0.05$ ; we therefore fail to reject the null hypothesis at this significance level.

The ideas illustrated by this example can now be generalized. As with previous discussions in Chapter 14, we organize the material according to the status of the population standard deviation,  $\sigma$ , because whether it is known or not determines what sampling distribution — and hence test statistic — is appropriate.

### 15.3.1 $\sigma$ Known; the “z-test”

*Problem:* The random variable,  $X$ , possesses a distribution,  $N(\mu, \sigma^2)$ , with unknown value,  $\mu$ , but known  $\sigma$ ; a random sample,  $X_1, X_2, \dots, X_n$ , is drawn from this normal population from which a sample average,  $\bar{X}$ , can be computed; a specific value,  $\mu_0$ , is hypothesized for the true population parameter; and it is desired to test whether the sample indeed came from such a population.

*The Hypotheses:* In testing such a hypothesis — concerning a single mean of a normal population with known standard deviation,  $\sigma$  — the null hypothesis is typically:

$$H_0 : \mu = \mu_0 \quad (15.16)$$

where  $\mu_0$  is the specific value postulated for the population mean (e.g. 75.5 used in the previous illustration). There are three possible alternative hypotheses:

$$H_a : \mu < \mu_0 \quad (15.17)$$

for the *lower-tailed*, one-sided (or one-tailed) alternative hypothesis; or

$$H_a : \mu > \mu_0 \quad (15.18)$$

for the *upper-tailed*, one-sided (or one-tailed) alternative hypothesis; or, finally, as illustrated above,

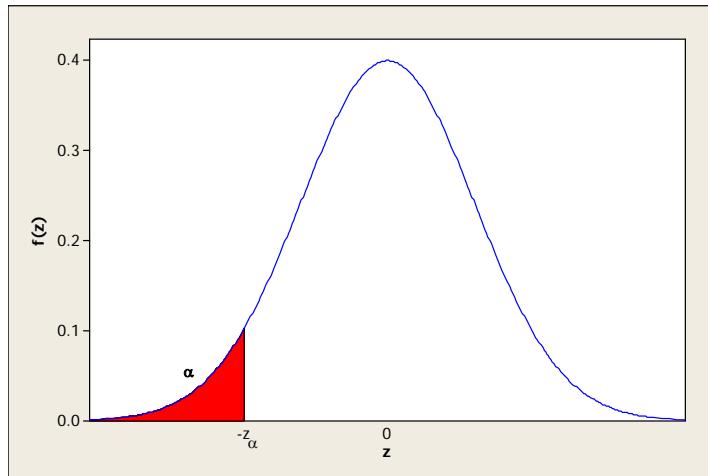
$$H_a : \mu \neq \mu_0 \quad (15.19)$$

for the two-sided (or two-tailed) alternative.

*Assumptions:* The underlying distribution in question is Gaussian, with known standard deviation,  $\sigma$ , implying that the sampling distribution of  $\bar{X}$  is also Gaussian, with mean,  $\mu_0$ , and variance,  $\sigma^2/n$ , if  $H_0$  is true. Hence, the random variable  $Z = (\bar{X} - \mu_0)/(\sigma/\sqrt{n})$  has a standard normal distribution,  $N(0, 1)$ .

*Test statistic:* The appropriate test statistic is therefore

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \quad (15.20)$$



**FIGURE 15.3:** The standard normal variate  $z = -z_\alpha$  with tail area probability  $\alpha$ . The shaded portion is the rejection region for a lower-tailed test,  $H_a : \mu < \mu_0$

The specific value obtained for a particular sample data average,  $\bar{x}$ , is sometimes called the “*z-score*” of the sample data.

*Critical/Rejection Regions:*

- (i) For lower-tailed tests (with  $H_a : \mu < \mu_0$ ), reject  $H_0$  in favor of  $H_a$  if:

$$z < -z_\alpha \quad (15.21)$$

where  $z_\alpha$  is the value of the standard normal variate,  $z$ , with a tail area probability of  $\alpha$ ; i.e.,  $P(z > z_\alpha) = \alpha$ . By symmetry,  $P(z < -z_\alpha) = P(z > z_\alpha) = \alpha$ , as shown in Fig 15.3. The rationale is that if  $\mu = \mu_0$  is true, then it is highly unlikely that  $z$  will be less than  $-z_\alpha$  by pure chance alone; it is more likely that  $\mu$  is systematically less than  $\mu_0$  if  $z$  is less than  $-z_\alpha$ .

(ii) For upper-tailed tests (with  $H_a : \mu > \mu_0$ ), reject  $H_0$  in favor of  $H_a$  if (see Fig 15.4):

$$z > z_\alpha \quad (15.22)$$

- (iii) For two-sided tests, (with  $H_a : \mu \neq \mu_0$ ), reject  $H_0$  in favor of  $H_a$  if:

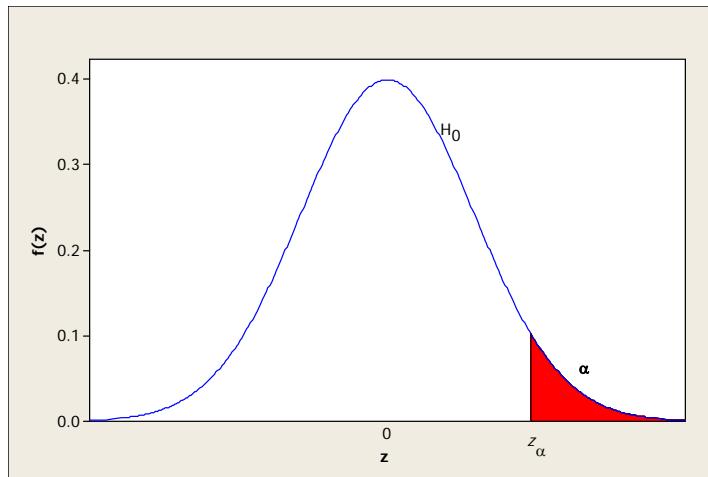
$$z < -z_{\alpha/2} \text{ or } z > z_{\alpha/2} \quad (15.23)$$

for the same reasons as above, because if  $H_0$  is true, then

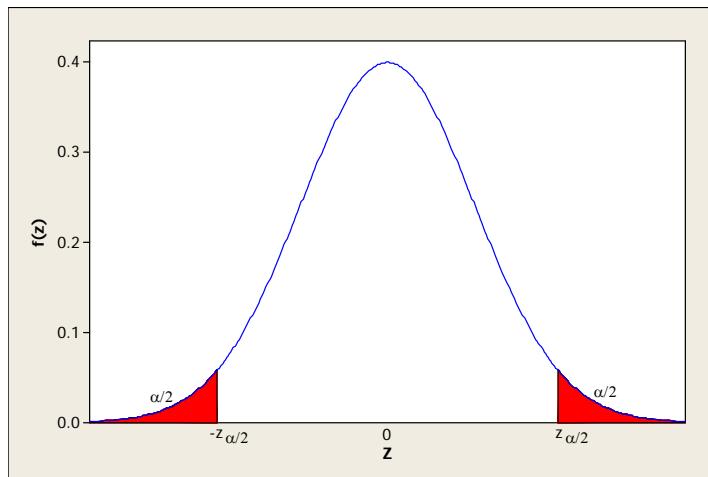
$$P(z < -z_{\alpha/2} \text{ or } z > z_{\alpha/2}) = \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha \quad (15.24)$$

as illustrated in Fig 15.5.

Tests of this type are known as “*z-tests*” because of the test statistic (and sampling distribution) upon which the test is based. Therefore,



**FIGURE 15.4:** The standard normal variate  $z = z_\alpha$  with tail area probability  $\alpha$ . The shaded portion is the rejection region for an upper-tailed test,  $H_a : \mu > \mu_0$



**FIGURE 15.5:** Symmetric standard normal variates  $z = z_{\alpha/2}$  and  $z = -z_{\alpha/2}$  with identical tail area probabilities  $\alpha/2$ . The shaded portions show the rejection regions for a two-sided test,  $H_a : \mu \neq \mu_0$

**TABLE 15.2:** Summary of  $H_0$  rejection conditions for the one-sample  $z$ -test

| Testing Against        | For general $\alpha$<br>Reject $H_0$ if:        | For $\alpha = 0.05$<br>Reject $H_0$ if: |
|------------------------|---|---|
| $H_a : \mu < \mu_0$    | $z < -z_\alpha$                                 | $z < -1.65$                             |
| $H_a : \mu > \mu_0$    | $z > z_\alpha$                                  | $z < 1.65$                              |
| $H_a : \mu \neq \mu_0$ | $z < -z_{\alpha/2}$<br>or<br>$z > z_{\alpha/2}$ | $z < -1.96$<br>or<br>$z > 1.96$         |

The *one-sample*  $z$ -test is a hypothesis test concerning the mean of a normal population where the population standard deviation,  $\sigma$ , is specified.

The key facts about the  $z$ -test for testing  $H_0 : \mu = \mu_0$  are summarized in Table 15.2.

The following two examples illustrate the application of the “ $z$ -test.”

**Example 15.2: CHARACTERIZING YIELD FROM PROCESS B**

Formulate and test (at the significance level of  $\alpha = 0.05$ ) the hypothesis implied by the second half of the statement given at the beginning of this chapter about the mean yield of process B, i.e., that  $Y_B \sim N(72.5, 2.5^2)$ . Use the data given in Chapter 1 and analyzed previously in various Chapter 14 examples.

**Solution:**

In this case, as with the  $Y_A$  illustration used to start this section, the hypotheses to be tested are:

$$\begin{aligned} H_0 : \mu_B &= 72.5 \\ H_a : \mu_B &\neq 72.5 \end{aligned} \quad (15.25)$$

a two-sided test. From the supplied data, we obtain  $\bar{y}_B = 72.47$ ; and since the population standard deviation,  $\sigma_B$ , is given as 2.5, the specific value,  $z$ , of the appropriate test statistic,  $Z$  (the “ $z$ -score”), from Eq (15.20), is:

$$z = \frac{72.47 - 72.50}{2.5/\sqrt{50}} = -0.084 \quad (15.26)$$

For this two-sided test, the critical value to the right,  $z_{\alpha/2}$ , for  $\alpha = 0.05$ , is:

$$z_{0.025} = 1.96 \quad (15.27)$$

so that the critical/rejection region,  $R_C$ , is  $z > 1.96$  to the right, in conjunction with  $z < -1.96$  to the left, by symmetry (recall Eq (15.14)).

And now, because the specific value  $z = -0.084$  does not lie in the critical/rejection region, we find no evidence to reject  $H_0$  in favor of the alternative. We conclude therefore that  $Y_B$  is very likely well-characterized by the postulated distribution.

We could also compute the  $p$ -value associated with this test statistic

$$P(z < -0.084 \text{ or } z > 0.084) = P(|z| > 0.084) = 0.933 \quad (15.28)$$

with the following implication: if  $H_0$  is true, the probability of observing, by pure chance alone, the actually observed sample average,  $\bar{y}_B = 72.47$ , or something “more extreme” (further away from the hypothesized mean of 72.50) is 0.933. Thus, there is no evidence to support rejecting  $H_0$ . Furthermore, since this  $p$ -value is much higher than the prescribed significance level,  $\alpha = 0.05$ , we cannot reject the null hypothesis at this significance level.

### Using MINITAB

It is instructive to walk through the typical procedure for carrying out such  $z$ -tests using computer software, in this case, MINITAB. From the MINITAB drop down menu, the sequence **Stat > Basic Statistics > 1-Sample Z** opens a dialog box that allows the user to carry out the analysis either using data already stored in MINITAB worksheet columns or from summarized data. Since we already have summarized data, upon selecting the “Summarized data” option, one enters 50 into the “Sample size:” dialog box, 72.47 into the “Mean:” box, and 2.5 into the “Standard deviation:” box; and upon selecting the “Perform hypothesis test” option, one enters 72.5 for the “Hypothesized mean.” The **OPTIONS** button allows the user to select the confidence level (the default is 95.0) and the “Alternative” for  $H_a$ : with the 3 available options displayed as “less than,” “not equal,” and “greater than.” The MINITAB results are displayed as follows:

| <b>One-Sample Z</b>                  |        |         |                  |       |       |  |
|--------------------------------------|--------|---------|------------------|-------|-------|--|
| Test of mu = 72.5 vs not = 72.5      |        |         |                  |       |       |  |
| The assumed standard deviation = 2.5 |        |         |                  |       |       |  |
| N                                    | Mean   | SE Mean | 95% CI           | Z     | P     |  |
| 50                                   | 72.470 | 0.354   | (71.777, 73.163) | -0.08 | 0.932 |  |

This output links hypothesis testing directly with estimation (as we anticipated in Chapter 14, and as we discuss further below) as follows: “SE Mean” is the standard error of the mean ( $\sigma/\sqrt{n}$ ) from which the 95% confidence interval (shown in the MINITAB output as “95% CI”) is obtained as (71.777, 73.163). Observe that the hypothesized mean, 72.5, is contained within this interval, with the implication that, since, at the 95% confidence level, the estimated average encompasses the hypothesized mean, we have no reason to reject  $H_0$  at the significance level of 0.05. The  $z$  statistic computed

by MINITAB is precisely what we had obtained in the example; the same is true of the  $p$ -value.

The results of this example (and the ones obtained earlier for  $Y_A$ ) may now be used to answer the first question raised at the beginning of this chapter (and in Chapter 1) regarding whether or not  $Y_A$  and  $Y_B$  consistently exceed 74.5.

The random variable,  $Y_A$ , has now been completely characterized by the Gaussian distribution,  $N(75.5, 1.5^2)$ , and  $Y_B$  by  $N(72.5, 2.5^2)$ . From these probability distributions, we are able to compute the following probabilities (using MINTAB):

$$P(Y_A > 74.5) = 1 - P(Y_A < 74.5) = 0.748 \quad (15.29)$$

$$P(Y_B > 74.5) = 1 - P(Y_B < 74.5) = 0.212 \quad (15.30)$$

The sequence for calculating cumulative probabilities is as follows: **Calc > Prob Dist > Normal**, which opens a dialog box for entering the desired parameters: (i) from the choices “Probability density,” “Cumulative Probability” and “Inverse Cumulative Probability,” one selects the second one; “Mean” is specified as 75.5 for the  $Y_A$  distribution, “Standard deviation” is specified as 1.5; and upon entering the input constant as 74.5, MINITAB returns the following results:

| Cumulative Distribution Function                            |          |
|---|----------|
| <u>Normal with mean = 75.5 and standard deviation = 1.5</u> |          |
| x   | P(X≤x)   |
| 74.5  | 0.252493 |

from which the required probability is obtained as  $1 - 0.252 = 0.748$ . Repeating the procedure for  $Y_B$ , with “Mean” specified as 72.5 and “Standard deviation” as 2.5 produces the result shown in Eq (15.30).

The implication of these results is that process A yields will exceed 74.5% around three-quarters of the time, whereas with the incumbent process B, exceeding yields of 74.5% will occur only one-fifths of the time. If profitability is related to yields that exceed 74.5% consistently, then process A will be roughly 3.5 times more profitable than the incumbent process B.

This next example illustrates how, in solving practical problems, “intuitive” reasoning without the objectivity of a formal hypothesis test can be misleading.

#### **Example 15.3: CHARACTERIZING “FAST-ACTING” RAT POISON**

The scientists at the ACME rat poison laboratories, who have been working non-stop to develop a new “fast-acting” formulation that will break the “thousand-second” barrier, appear to be on the verge of a breakthrough. Their target is a product that will kill rats within 1000

secs, on average, with a standard deviation of 125 secs. Experimental tests conducted in an affiliated toxicology laboratory in which pellets were made with a newly developed formulation and administered to 64 rats (selected at random from an essentially identical population). The results showed an average “acting time,”  $\bar{x} = 1028$  secs. The ACME scientists, anxious to declare a breakthrough, were preparing to approach management immediately to argue that the observed excess 28 secs, when compared to the stipulated standard deviation of 125 seconds, is “small and insignificant.” The group statistician, in an attempt to present an objective, statistically-sound argument, recommended instead that a hypothesis test should first be carried out to rule out the possibility that the mean “acting time” is still greater than 1000 secs. Assuming that the “acting time” measurements are normally distributed, carry out an appropriate hypothesis test and, at the significance level of  $\alpha = 0.05$ , make an informed recommendation regarding the tested rat poison’s “acting time.”

**Solution:**

For this problem, the null and alternative hypotheses are:

$$\begin{aligned} H_0 : \mu &= 1000 \\ H_a : \mu &> 1000 \end{aligned} \quad (15.31)$$

The alternative has been chosen this way because the concern is that the acting time may still be greater than 1000 secs. As a result of the normality assumption, and the fact that  $\sigma$  is specified as 125, the required test is the  $z$ -test, where the specific  $z$ -score, from Eq (15.20), in this case is:

$$z = \frac{1028 - 1000}{125/\sqrt{64}} = 1.792 \quad (15.32)$$

The critical value,  $z_\alpha$ , for  $\alpha = 0.05$  for this upper-tailed one-sided test is:

$$z_{0.05} = 1.65, \quad (15.33)$$

obtained from MINITAB using the inverse cumulative probability feature for the standard normal probability distribution with tail area probability 0.05, i.e.,

$$P(Z > 1.65) = 0.05 \quad (15.34)$$

Thus, the rejection region,  $R_C$ , is  $z > 1.65$ . And now, because  $z = 1.78$  falls into the rejection region, the decision is to reject the null hypothesis at the 5% level. Alternatively, the  $p$ -value associated with this test statistic can be obtained (also from MINITAB, using the cumulative probability feature) as:

$$P(z > 1.792) = 0.037, \quad (15.35)$$

implying that if  $H_0$  is true, the probability of observing, by pure chance alone, the actually observed sample average, 1028 secs, or something higher, is so small that we are inclined to believe that  $H_0$  is unlikely to be true. Observe that this  $p$ -value is lower than the specified significance level of  $\alpha = 0.05$ .

Thus, from these equivalent perspectives, the conclusion is that the experimental evidence *does not* support the ACME scientists premature declaration of a breakthrough; the observed excess 28 secs, in fact, appears to be significant at the  $\alpha = 0.05$  significance level.

Using the procedure illustrated previously, the MINITAB results for this problem are displayed as follows:

| <b>One-Sample Z</b>                  |        |         |                 |      |       |  |
|--------------------------------------|--------|---------|-----------------|------|-------|--|
| Test of $\mu = 1000$ vs $> 1000$     |        |         |                 |      |       |  |
| The assumed standard deviation = 125 |        |         |                 |      |       |  |
| N                                    | Mean   | SE Mean | 95% Lower Bound | Z    | P     |  |
| 64                                   | 1028.0 | 15.6    | 1002.3          | 1.79 | 0.037 |  |

Observe that the *z*- and *p*-values agree with what we had obtained earlier; furthermore, the additional entries, “SE Mean,” for the standard error of the mean, 15.6, and the 95% lower bound on the estimate for the mean, 1002.3, link this hypothesis test to interval estimation. This connection will be explored more fully later in this section; for now, we note simply that the 95% lower bound on the estimate for the mean, 1002.3, lies entirely to the right of the hypothesized mean value of 1000. The implication is that, at the 95% confidence level, it is more likely that the true mean is *higher* than the value hypothesized; we are therefore more inclined to reject the null hypothesis in favor of the alternative, at the significance level 0.05.

### 15.3.2 $\sigma$ Unknown; the “t-test”

When the population standard deviation,  $\sigma$ , is unknown, the sample standard deviation,  $s$ , will have to be substituted for it. In this case, one of two things can happen:

1. If the sample size is sufficiently large (for example,  $n > 30$ ),  $s$  is usually considered to be a good enough approximation to  $\sigma$ , that the *z*-test can be applied, treating  $s$  as equal to  $\sigma$ .
2. When the sample size is small, substituting  $s$  for  $\sigma$  changes the test statistic and the corresponding test, as we now discuss.

For small sample sizes, when  $S$  is substituted for  $\sigma$ , the appropriate test statistic, becomes

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \quad (15.36)$$

which, from our discussion of sampling distributions, is known to possess a Student’s *t*-distribution, with  $\nu = n - 1$  degrees of freedom. This is the “small sample size” equivalent of Eq (15.20).

Once more, because of the test statistic, and the sampling distribution upon which the test is based, this test is known as a “*t*-test.” Therefore,

**TABLE 15.3:** Summary of  $H_0$   
rejection conditions for the  
one-sample  $t$ -test

| Testing Against        | For general $\alpha$<br>Reject $H_0$ if:                                    |
|------------------------|---|
| $H_a : \mu < \mu_0$    | $t < -t_\alpha(\nu)$  |
| $H_a : \mu > \mu_0$    | $t > t_\alpha(\nu)$   |
| $H_a : \mu \neq \mu_0$ | $t < -t_{\alpha/2}(\nu)$ or<br>$t > t_{\alpha/2}(\nu)$<br>( $\nu = n - 1$ ) |

The *one-sample t*-test is a hypothesis test concerning the mean of a normal population when the population standard deviation,  $\sigma$ , is unknown, and the sample size is small.

The  $t$ -test is therefore the same as the  $z$ -test but with the sample standard deviation,  $s$ , used in place of the unknown  $\sigma$ ; it uses the  $t$ -distribution (with the appropriate degrees of freedom) in place of the standard normal distribution of the  $z$ -test. The relevant facts about the  $t$ -test for testing  $H_0 : \mu = \mu_0$  are summarized in Table 15.3, the equivalent of Table 15.2 shown earlier. The specific test statistic,  $t$ , is determined by introducing sample data into Eq 15.36. Unlike the  $z$ -test, even after specifying  $\alpha$ , we are unable to determine the specific critical/rejection region because these values depend on the degrees of freedom (i.e., the sample size). The following example illustrates how to conduct a one-sample  $t$ -test.

**Example 15.4: HYPOTHESES TESTING REGARDING ENGINEERING TRAINING PROGRAMS**

Assume that the test results shown in Example 15.1 are random samples from normal populations. (1) At a significance level of  $\alpha = 0.05$ , test the hypothesis that the mean score for trainees using method A is  $\mu_A = 75$ , versus the alternative that it is less than 75. (2) Also, at the same significance level, test the hypothesis that the mean score for trainees using method B is  $\mu_B = 75$ , versus the alternative that it is not.

**Solution:**

(1) The first thing to note is that the population standard deviations are not specified; and since the sample size of 10 for each data set is small, the appropriate test is a one-sample  $t$ -test. The null and alternative hypotheses for the first problem are:

$$\begin{aligned} H_0 : \mu_A &= 75.0 \\ H_a : \mu_A &< 75.0 \end{aligned} \tag{15.37}$$

The sample average is obtained from the supplied data as  $\bar{x}_A = 69.0$ ,

with a sample standard deviation,  $s_A = 4.85$ ; the specific  $T$  statistic value is thus obtained as:

$$t = \frac{69.0 - 75.0}{4.85/\sqrt{10}} = -3.91 \quad (15.38)$$

Because this is a lower-tailed, one-sided test, the critical value,  $-t_{0.05}(9)$ , is obtained as  $-1.833$  (using MINITAB's inverse cumulative probability feature, for the  $t$ -distribution with 9 degrees of freedom). The rejection region,  $R_C$ , is therefore  $t < -1.833$ . Observe that the specific  $t$ -value for this test lies well within this rejection region; we therefore reject the null hypothesis in favor of the alternative, at the significance level 0.05.

Of course, we could also compute the  $p$ -value associated with this particular test statistic; and from the  $t$ -distribution with 9 degrees of freedom we obtain,

$$P(T(9) < -3.91) = 0.002 \quad (15.39)$$

using MINITAB's cumulative probability feature. The implication here is that the probability of observing a difference as large, or larger, between the postulated mean (75) and actual sample average (69), if  $H_0$  is true, is so very low (0.002) that it is more likely that the alternative is true; that the sample average is more likely to have come from a distribution whose mean is less than 75. Equivalently since this  $p$ -value is less than the significance level 0.05, we reject  $H_0$  at this significance level.

(2) The hypotheses to be tested in this case are:

$$\begin{aligned} H_0 : \mu_B &= 75.0 \\ H_a : \mu_B &\neq 75.0 \end{aligned} \quad (15.40)$$

From the supplied data, the sample average and standard deviation are obtained respectively as  $\bar{x}_B = 74.0$ , and  $s_B = 5.40$ , so that the specific value for the  $T$  statistic is:

$$t = \frac{74 - 75.0}{5.40/\sqrt{10}} = -0.59 \quad (15.41)$$

Since this is a two-tailed test, the critical values,  $t_{0.025}(9)$  and its mirror image  $-t_{0.025}(9)$ , are obtained from MINITAB as:  $-2.26$  and  $2.26$  implying that the critical/rejection region,  $R_C$ , in this case is  $t < -2.26$  or  $t > 2.26$ . But the specific value for the  $t$ -statistic ( $-0.59$ ) does not lie in this region; we therefore *do not* reject  $H_0$  at the significance level 0.05.

The associated  $p$ -value, obtained from a  $t$ -distribution with 9 degrees of freedom, is:

$$P(t(9) < -0.59 \text{ or } t(9) > 0.59) = P(|t(9)| > 0.59) = 0.572 \quad (15.42)$$

with the implication that we do not reject the null hypothesis, either on the basis of the  $p$ -value, or else at the 0.05 significance level, since  $p = 0.572$  is larger than 0.05.

Thus, observe that with these two *t*-tests, we have established, at a significance level of 0.05, that the mean score obtained by trainees using method A is less than 75 while the mean score for trainees using method B is essentially equal to 75. We can, of course, infer from here that this means that method B must be more effective. But there are more direct methods for carrying out tests to compare two means directly, which will be considered shortly.

### Using MINITAB

MINITAB can be used to carry out these *t*-tests directly (without having to compute, by ourselves, first the test statistic and then the critical region, etc). After entering the data into separate columns, “Method A” and “Method B” in a MINITAB worksheet, for the first problem, the sequence **Stat > Basic Statistics > 1-Sample t** from the MINITAB drop down menu opens a dialog box where one selects the column containing the data, (“Method A”); and upon selecting the “Perform hypothesis test” option, one enters the appropriate value for the “Hypothesized mean” (75) and with the **OPTIONS** button one selects the desired “Alternative” for  $H_a$  (less than) along with the default confidence level (95.0).

MINITAB provides three self-explanatory graphical options: “Histogram of data;” “Individual value plot;” and “Boxplot of data.” Our discussion in Chapter 12 about graphical plots for small sample data sets recommends that, with  $n = 10$  in this case, the box plot is more reasonable than the histogram for this example.

The resulting MINITAB outputs are displayed as follows:

#### One-Sample T: Method A

Test of mu = 75 vs < 75

| Variable | N  | Mean  | StDev | SE Mean | 95% Upper |  | T     | P     |
|----------|----|-------|-------|---------|-----------|--|-------|-------|
|          |    |       |       |         | Bound     |  |       |       |
| Method A | 10 | 69.00 | 4.85  | 1.53    | 71.81     |  | -3.91 | 0.002 |

The box-plot along with the 95% confidence interval estimate and the hypothesized mean  $H_0 = 75$  are shown in Fig 15.6. The conclusion to reject the null hypothesis in favor of the alternative is clear.

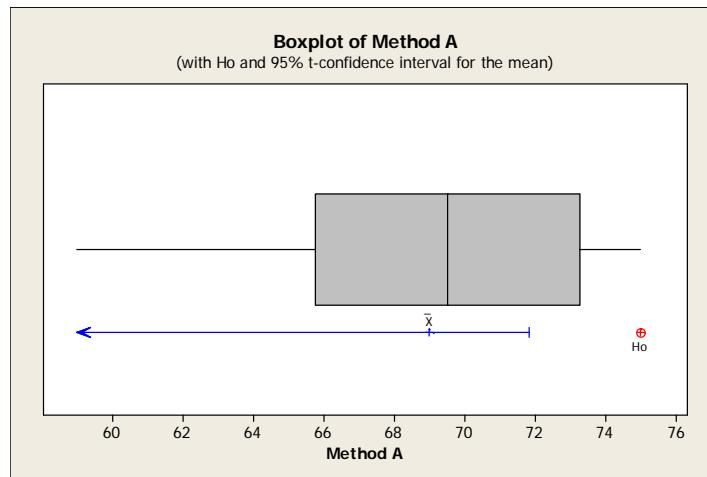
In dealing with the second problem regarding Method B, we follow the same procedure, selecting data in the “Method B” column, but this time, the “Alternative” is selected as “not equal.” The MINITAB results are displayed as follows:

#### One-Sample T: Method B

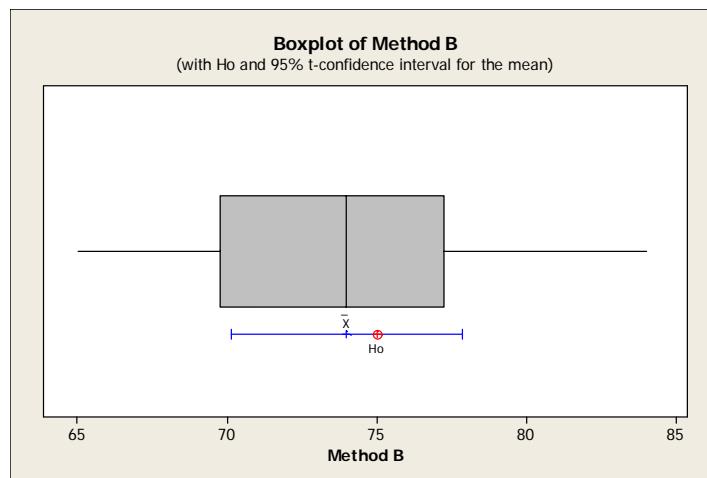
Test of mu = 75 vs not = 75

| Variable | N  | Mean  | StDev | SE Mean | 95% CI         | T     | P     |
|----------|----|-------|-------|---------|----------------|-------|-------|
| Method B | 10 | 74.00 | 5.40  | 1.71    | (70.14, 77.86) | -0.59 | 0.572 |

The box-plot along with the 95% confidence interval for the mean and the hypothesized mean  $H_0 = 75$  are shown in Fig 15.7.



**FIGURE 15.6:** Box plot for Method A scores including the null hypothesis mean,  $H_0 : \mu = 75$ , shown along with the sample average,  $\bar{x}$ , and the 95% confidence interval based on the  $t$ -distribution with 9 degrees of freedom. Note how the upper bound of the 95% confidence interval lies to the left of, and does not touch, the postulated  $H_0$  value



**FIGURE 15.7:** Box plot for Method B scores including the null hypothesis mean,  $H_0 : \mu = 75$ , shown along with the sample average,  $\bar{x}$ , and the 95% confidence interval based on the  $t$ -distribution with 9 degrees of freedom. Note how the the 95% confidence interval includes the postulated  $H_0$  value

### 15.3.3 Confidence Intervals and Hypothesis Tests

Interval estimation techniques discussed in Chapter 14 produced estimates for the parameter  $\theta$  in the form of an interval, ( $u_L < \theta < u_R$ ), that is expected to contain the unknown parameter with probability  $(1 - \alpha)$ ; it is therefore known as the  $(1 - \alpha) \times 100\%$  confidence interval.

Now, observe first from the definition of the critical/rejection region,  $R_C$ , given above, first for a two-tailed test, that at the significance level,  $\alpha$ ,  $R_C$  is precisely complementary to the  $(1 - \alpha) \times 100\%$  confidence interval for the estimated parameter. The implication therefore is as follows: if the postulated population parameter (say  $\theta_0$ ) falls *outside* the  $(1 - \alpha) \times 100\%$  confidence interval estimated from sample data, (i.e., the postulated value is higher than the upper bound to the right, or lower than the lower bound to the left) this triggers the rejection of  $H_0$ , that  $\theta = \theta_0$ , at the significance level of  $\alpha$ , in favor of the alternative  $H_a$ , that  $\theta \neq \theta_0$ . Conversely, if the postulated  $\theta_0$  falls within the  $(1 - \alpha) \times 100\%$  confidence interval, we will fail to reject  $H_0$ . This is illustrated in Example 15.2 for the mean yield of process B. The 95% confidence interval was obtained as (70.74, 74.20), which fully encompasses the hypothesized mean value of 72.5; hence we do not reject  $H_0$  at the 0.05 significance level. Similarly, in part 2 of Example 15.4, the 95% confidence interval on the average method B score was obtained as (70.14, 77.86); and with the hypothesized mean, 75, lying entirely in this interval (as shown graphically in Fig 15.7). Once again, we find no evidence to reject  $H_0$  at the 0.05 significance level.

For an upper-tailed test (with  $H_a$  defined as  $H_a : \theta > \theta_0$ ), it is the *lower bound* of the  $(1 - \alpha) \times 100\%$  confidence interval that is now of interest. Observe that if the hypothesized value,  $\theta_0$ , is to the *left* of this lower bound (i.e., it is lower than the lowest value of the  $(1 - \alpha) \times 100\%$  confidence interval), the implication is twofold: (i) the computed estimate falls in the rejection region; and, equivalently, (ii) value estimated from data is larger than the hypothesized value — both of which support the rejection of  $H_0$  in favor of  $H_a$ , at the significance level of  $\alpha$ . This is illustrated in Example 15.3 where the lower bound of the estimated “acting time” for the rat poison was obtained (from MINITAB) as 1002.3 secs, whereas the postulated mean is 1000.  $H_0$  is therefore rejected at the 0.05 significance level in favor of  $H_a$ , that the mean value is higher. On the other hand, if the hypothesized value,  $\theta_0$ , is to the *right* of this lower bound, there will be no support for rejecting  $H_0$  at the 0.05 significance level.

The reverse is true for the lower-tailed test with  $H_a : \theta < \theta_0$ . The *upper bound* of the  $(1 - \alpha) \times 100\%$  confidence interval is of interest; and if the hypothesized value,  $\theta_0$ , is to the *right* of this upper bound (i.e., it is larger than the largest value of the  $(1 - \alpha) \times 100\%$  confidence interval), this hypothesized value would have fallen into the rejection region. Because this indicates that the value estimated from data is smaller than the hypothesized value, the evidence supports the rejection of  $H_0$  in favor of  $H_a$ , at the 0.05 significance

level. Again, this is illustrated in part 1 of Example 15.4. The upper bound of the 95% confidence interval on the average method A score was obtained as 71.81, which is lower than the postulated average of 75, thereby triggering the rejection of  $H_0$  in favor of  $H_a$ , at the 0.05 significance level (see Fig 15.6). Conversely, when the hypothesized value,  $\theta_0$ , is to the *left* of this upper bound, we will fail to reject  $H_0$  at the 0.05 significance level.

## 15.4 Concerning Two Normal Population Means

The problem of interest involves two distinct and *mutually independent* normal populations, with respective unknown means  $\mu_1$  and  $\mu_2$ . In general we are interested in making inference about the difference between these two means, i.e.,

$$\mu_1 - \mu_2 = \delta \quad (15.43)$$

The typical starting point is the null hypothesis,

$$H_0 : \mu_1 - \mu_2 = \delta_0 \quad (15.44)$$

when the difference between the two population means is postulated as some value  $\delta_0$ , and the hypothesis is to be tested against the usual triplet of possible alternatives:

$$\text{Lower-tailed } H_a : \mu_1 - \mu_2 < \delta_0 \quad (15.45)$$

$$\text{Upper-tailed } H_a : \mu_1 - \mu_2 > \delta_0 \quad (15.46)$$

$$\text{Two-tailed } H_a : \mu_1 - \mu_2 \neq \delta_0 \quad (15.47)$$

In particular, specifying  $\delta_0 = 0$  constitutes a test of equality of the two means; but  $\delta_0$  does not necessarily have to be zero, allowing us to test the difference against any arbitrary postulated value.

As with tests of single population means, this test will be based on the difference between two random sample means,  $\bar{X}_1$  from population 1, and  $\bar{X}_2$  from population 2. These tests are therefore known as “two-sample” tests; and, as usual, the specific test to be employed for any problem depends on what additional information is available about each population’s standard deviation.

### 15.4.1 Population Standard Deviations Known

When the population standard deviations,  $\sigma_1$  and  $\sigma_2$  are known, we recall (from the discussion in Chapter 14 on interval estimation of the difference of

**TABLE 15.4:** Summary of  $H_0$  rejection conditions for the two-sample  $z$ -test

| Testing Against                     | For general $\alpha$<br>Reject $H_0$ if:     | For $\alpha = 0.05$<br>Reject $H_0$ if: |
|-------------------------------------|--|---|
| $H_a : \mu_1 - \mu_2 < \delta_0$    | $z < -z_\alpha$                              | $z < -1.65$                             |
| $H_a : \mu_1 - \mu_2 > \delta_0$    | $z > z_\alpha$                               | $z < 1.65$                              |
| $H_a : \mu_1 - \mu_2 \neq \delta_0$ | $z < -z_{\alpha/2}$ or<br>$z > z_{\alpha/2}$ | $z < -1.96$ or<br>$z > 1.96$            |

two normal population means) that the test statistic:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1) \quad (15.48)$$

where  $n_1$  and  $n_2$  are the sizes of the samples drawn from populations 1 and 2 respectively. This fact arises from the result established in Chapter 14 for the sampling distribution of  $\bar{D} = \bar{X}_1 - \bar{X}_2$  as  $N(\delta, v^2)$ , with  $\delta$  as defined in Eq (18.10), and

$$v^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \quad (15.49)$$

Tests based on this statistic are known as “two-sample  $z$ -tests,” and as with previous tests, the specific results for testing  $H_0 : \mu_1 - \mu_2 = \delta_0$  are summarized in Table 15.4.

Let us illustrate the application of this test with the following example.

**Example 15.5: COMPARISON OF SPECIALTY AUXILIARY BACKUP LAB BATTERY LIFETIMES**

A company that manufactures specialty batteries used as auxiliary backups for sensitive laboratory equipments in need of constant power supplies claims that its new prototype, brand A, has a longer lifetime (under constant use) than the industry-leading brand B, and at the same cost. Using accepted industry protocol, a series of tests carried out in an independent laboratory produced the following results: For brand A: sample size,  $n_1 = 40$ ; average lifetime,  $\bar{x}_1 = 647$  hrs; with a population standard deviation given as  $\sigma_1 = 27$  hrs. The corresponding results for brand B are  $n_2 = 40$ ;  $\bar{x}_2 = 638$ ;  $\sigma_2 = 31$ . Determine, at the 5% level, if there is a significant difference between the observed mean lifetimes.

**Solution:**

Observe that in this case,  $\delta_0 = 0$ , i.e., the null hypothesis is that the two means are equal; the alternative is that  $\mu_1 > \mu_2$ , so that the hypotheses are formulated as:

$$\begin{aligned} H_0 &: \mu_1 - \mu_2 = 0 \\ H_a &: \mu_1 - \mu_2 > 0 \end{aligned} \quad (15.50)$$

The specific test statistic obtained from the experimental data is:

$$z = \frac{(647 - 638) - 0}{\sqrt{\frac{27^2}{40} + \frac{31^2}{40}}} = 1.38 \quad (15.51)$$

For this one-tailed test, the critical value,  $z_{0.05}$ , is 1.65; and now, since the computed  $z$ -score is not greater than 1.65, we cannot reject the null hypothesis. There is therefore insufficient evidence to support the rejection of  $H_0$  in favor of  $H_a$ , at the 5% significance level.

Alternatively, we could compute the  $p$ -value and obtain:

$$\begin{aligned} p = P(Z > 1.38) &= 1 - P(Z < 1.38) \\ &= 1 - 0.916 = 0.084 \end{aligned} \quad (15.52)$$

Once again, since this  $p$ -value is greater than 0.05, we cannot reject  $H_0$  in favor of  $H_a$ , at the 5% significance level. (However, observe that at the 0.1 significance level, we will reject  $H_0$  in favor of  $H_a$ , since the  $p$ -value is less than 0.1.)

#### 15.4.2 Population Standard Deviations Unknown

In most practical cases, it is rare that the two population standard deviations are known. Under these circumstances, we are able to identify three distinct cases requiring different approaches:

1.  $\sigma_1$  and  $\sigma_2$  unknown; large sample sizes  $n_1$  and  $n_2$ ;
2. Small sample sizes;  $\sigma_1$  and  $\sigma_2$  unknown, but equal (i.e.,  $\sigma_1 = \sigma_2$ );
3. Small sample sizes;  $\sigma_1$  and  $\sigma_2$  unknown, and unequal (i.e.,  $\sigma_1 \neq \sigma_2$ ).

As usual, under the first set of conditions, the sample standard deviations,  $s_1$  and  $s_2$ , are considered to be sufficiently good approximations to the respective unknown population parameters; they are then used in place of  $\sigma_1$  and  $\sigma_2$  in carrying out the two-sample  $z$ -test as outlined above. Nothing more need be said about this case. We will concentrate on the remaining two cases where the sample sizes are considered to be small.

##### Equal standard deviations

When the two population standard deviations are considered as equal, the test statistic:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \sim t(\nu) \quad (15.53)$$

i.e., its sampling distribution is a  $t$ -distribution with  $\nu$  degrees of freedom, with

$$\nu = n_1 + n_2 - 2 \quad (15.54)$$

**TABLE 15.5:** Summary of  $H_0$  rejection conditions for the two-sample  $t$ -test

| Testing Against                     | For general $\alpha$<br>Reject $H_0$ if:  |
|-------------------------------------|---|
| $H_a : \mu_1 - \mu_2 < \delta_0$    | $t < -t_\alpha(\nu)$  |
| $H_a : \mu_1 - \mu_2 > \delta_0$    | $t > t_\alpha(\nu)$   |
| $H_a : \mu_1 - \mu_2 \neq \delta_0$ | $t < -t_{\alpha/2}(\nu)$ or<br>$t > t_{\alpha/2}(\nu)$<br>( $\nu = n_1 + n_2 - 2$ ) |

Here,  $S_p$  is the “pooled” sample standard deviation obtained as the positive square root of the pooled sample variance — a weighted average of the two sample variances:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}, \quad (15.55)$$

a reasonable estimate of the (equal) population variances based on the two sample variances.

From this test statistic and its sampling distribution, one can now carry out the “two-sample  $t$ -test,” and, once more, the specific results for testing  $H_0 : \mu_1 - \mu_2 = \delta_0$  against various alternatives are summarized in Table 15.5.

The following example illustrates these results.

**Example 15.6: HYPOTHESES TEST COMPARING EFFEC-TIVENESS OF ENGINEERING TRAINING PROGRAMS**

Revisit the problem in Example 15.1 and this time, at the 5% significance level, test the claim that Method B is more effective. Assume that the scores shown in Example 1 come from normal populations with potentially different means, but equal variances.

**Solution:**

In this case, because the sample size is small for each data set, the appropriate test is a two-sample  $t$ -test, with equal variance; the hypotheses to be tested are:

$$\begin{aligned} H_0 &: \mu_A - \mu_B = 0 \\ H_a &: \mu_A - \mu_B < 0 \end{aligned} \quad (15.56)$$

Care must be taken in ensuring that  $H_a$  is specified properly. Since the claim is that Method B is more effective, if the difference in the means is specified in  $H_0$  as shown (with  $\mu_A$  first), then the appropriate  $H_a$  is as we have specified. (We are perfectly at liberty to formulate  $H_0$  differently, with  $\mu_B$  first, in which case the alternative hypothesis must change to  $H_a : \mu_B - \mu_A > 0$ .)

From the sample data, we obtain all the quantities required for computing the test statistic: the sample means,  $\bar{x}_A = 69.0$ ,  $\bar{x}_B = 74.0$ ; the sample standard deviations,  $s_A = 4.85$ ,  $s_B = 5.40$ ; so that the estimated pooled standard deviation is obtained as:

$$s_p = 5.13$$

with  $\nu = 18$ . To test the observed difference ( $d = 69.0 - 74.0 = -5.0$ ) against a hypothesized difference of  $\delta_0 = 0$  (i.e., equality of the means), we obtain the  $t$ -statistic as:

$$t = -2.18,$$

which is compared to the critical value for a  $t$ -distribution with 18 degrees of freedom,

$$-t_{0.05}(18) = -1.73.$$

And since  $t < -t_{0.05}(18)$ , we reject the null hypothesis in favor of the alternative, and conclude that, at the 5% significance level, the evidence in the data supports the claim that Method B is more effective.

Note also that the associated  $p$ -value, obtained from a  $t$  distribution with 18 degrees of freedom, is:

$$P(t(18) < -2.18) = 0.021 \quad (15.57)$$

which, by virtue of being less than 0.05 recommends rejection of  $H_0$  in favor of  $H_a$ , at the 5% significance level, as we already concluded above.

## Using MINITAB

This just-concluded example illustrates the “mechanics” of how to conduct a two-sample  $t$ -test “manually;” once the mechanics are understood, however, it is recommended to use computer programs such as MINITAB.

As noted before, once the data sets have been entered into separate columns “Method A” and “Method B” in a MINITAB worksheet (as was the case in the previous Example 15.4), the required sequence from the MINITAB drop down menu is: **Stat > Basic Statistics > 2-Sample t**, which opens a dialog box with self-explanatory options. Once the location of the relevant data are identified, the “Assume equal variance” box is selected in this case, and with the **OPTIONS** button, one selects the “Alternative” for  $H_a$  (“less than,” if the hypotheses are set up as we have done above), along with the default confidence level (95.0); one enters the value for hypothesized difference,  $\delta_0$ , in the “Test difference” box (0 in this case). The resulting MINITAB outputs for this problem are displayed as follows:

### Two-Sample T-Test and CI: Method A, Method B

#### Two-sample T for Method A vs Method B

|          | N  | Mean  | StDev | SE Mean |
|----------|----|-------|-------|---------|
| Method A | 10 | 69.00 | 4.85  | 1.5     |
| Method B | 10 | 74.00 | 5.40  | 1.7     |

```
Difference = mu (Method A) - mu (Method B)
Estimate for difference: -5.00
95% upper bound for difference: -1.02
T-Test of difference = 0 (vs <): T-Value = -2.18 P-Value = 0.021 DF
= 18
Both use Pooled StDev = 5.1316
```

### Unequal standard deviations

When  $\sigma_1 \neq \sigma_2$ , things become a bit more complicated, and a detailed discussion lies outside the intended scope of this book. Suffice it to say that under these circumstances, the universally recommended test statistic is  $\tilde{T}$  defined as:

$$\tilde{T} = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}, \quad (15.58)$$

which appears deceptively like Eq (15.53), with the very important difference that  $S_1$  and  $S_2$  have been reinstated individually in place of the pooled  $S_p$ . Of course, this expression is also reminiscent of the  $Z$  statistic in Eq (15.48), with  $S_1$  and  $S_2$  introduced in place of the population variances. However, unlike the other single variable cases where such a substitution transforms the standard normal sampling distribution to the  $t$ -distribution with the appropriate degrees of freedom, unfortunately, this time, this test statistic only has an *approximate* (not exact)  $t$ -distribution; and the degrees of freedom,  $\nu$ , accompanying this approximate  $t$ -distribution is defined by:

$$\nu = \tilde{n}_{12} - 2 \quad (15.59)$$

with  $\tilde{n}_{12}$  defined by the formidable-looking expression

$$\tilde{n}_{12} = \left\{ \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{\frac{(S_1^2/n_1)^2}{n_1+1} + \frac{(S_2^2/n_2)^2}{n_2+1}} \right\} \quad (15.60)$$

rounded to the nearest integer.

Under these conditions, the specific results for carrying out two-sample  $t$ -tests for testing  $H_0 : \mu_1 - \mu_2 = \delta_0$  against various alternatives are summarized in Table 15.5 but with  $\tilde{t}$  in place of the corresponding  $t$ -values, and using  $\nu$  given above in Eqs (15.59) and (15.60) for the degrees of freedom.

Although it is possible to carry out such two-sample  $t$ -tests “manually” by computing the required quantities on our own, it is highly recommended that such tests be carried out using computer programs such as MINITAB.

### Confidence Intervals and Two-Sample tests

The relationship between confidence intervals for the difference between two normal population means and the two-sample tests discussed above perfectly mirrors the earlier discussion concerning single means of a normal population. For the two-sided test, a  $(1 - \alpha) \times 100\%$  confidence interval estimate for

the difference between the two means that does not contain the hypothesized mean corresponds to a hypothesis test in which  $H_0$  is rejected, at the significance level of  $\alpha$ , in favor of the alternative that the computed difference is not equal to the hypothesized difference. Note that with a test of equality (in which case  $\delta_0$ , the hypothesized difference, is 0), rejection of  $H_0$  is tantamount to the  $(1 - \alpha) \times 100\%$  confidence interval for the difference not containing 0. On the contrary, an estimated  $(1 - \alpha) \times 100\%$  confidence interval that contains the hypothesized difference is equivalent to a two-sample test that must fail to reject  $H_0$ .

The corresponding arguments for the upper-tailed and lower-tailed tests follow precisely as presented earlier. For an upper-tailed test,  $(H_a : \delta > \delta_0)$ , a *lower bound* of the  $(1 - \alpha) \times 100\%$  confidence interval estimate of the difference,  $\delta$ , that is larger than the hypothesized difference,  $\delta_0$ , corresponds to a two-sample test in which  $H_0$  is rejected in favor of  $H_a$ , at the significance level of  $\alpha$ . Conversely, a *lower bound* of the confidence interval estimate of the difference,  $\delta$ , that is smaller than the hypothesized difference,  $\delta_0$ , corresponds to a test that will not reject  $H_0$ . The reverse is the case for the lower-tailed test  $(H_a : \delta < \delta_0)$ : when the upper bound of the  $(1 - \alpha) \times 100\%$  confidence interval estimate of  $\delta$  is smaller than  $\delta_0$ ,  $H_0$  is rejected in favor of  $H_a$ . When the upper bound of the  $(1 - \alpha) \times 100\%$  confidence interval estimate of  $\delta$  is larger than  $\delta_0$ ,  $H_0$  is *not* rejected.

### An Illustrative Example: The Yield Improvement Problem

The solution to the yield improvement problem first posed in Chapter 1, and revisited at the beginning of this chapter, will finally be completed in this illustrative example. In addition, the example also illustrates the use of MINITAB to carry out a two-sample  $t$ -test when population variances are not equal.

The following questions remain to be resolved: Is  $Y_A > Y_B$ , and if so, is  $Y_A - Y_B > 2$ ? Having already confirmed that the random variables,  $Y_A$  and  $Y_B$ , can be characterized reasonably well with Gaussian distributions,  $N(\mu_A, \sigma_A^2)$  and  $N(\mu_B, \sigma_B^2)$ , respectively, the supplied data may then be considered as being from normal distributions with *unequal* population variances. We will answer these two questions by carrying out appropriate two-sample  $t$ -tests.

Although the answer to the first of the two questions requires testing for the equality of  $\mu_A$  and  $\mu_B$  against the alternative that  $\mu_A > \mu_B$ , let us begin by first testing against  $\mu_A \neq \mu_B$ ; this establishes that the two distributions means are different. Later we will test against the alternative that  $\mu_A > \mu_B$ , and thereby go beyond the mere existence of a difference between the population means to establish which is larger. Finally, we proceed even one step further to establish not only which one is larger, but that it is larger by a value that exceeds a certain postulated value (in this case 2).

For the first test of basic equality, the hypothesized difference is clearly

$\delta_0 = 0$ , so that:

$$\begin{aligned} H_0 : \mu_A - \mu_B &= 0 \\ H_a : \mu_A &\neq \mu_B = 0 \end{aligned} \quad (15.61)$$

The procedure for using MINITAB is as follows: upon entering the data into separate  $Y_A$  and  $Y_B$  columns in a MINITAB worksheet, the required sequence from the MINITAB drop down menu is: **Stat > Basic Statistics > 2-Sample t**. In the opened dialog box, one simply selects the “Samples in different columns” option, identifies the columns corresponding to each data set, but this time, the “Assume equal variance” box must not be selected. With the **OPTIONS** button one selects the “Alternative” for  $H_a$  as “not equal,” along with the default confidence level (95.0); in the “Test difference” box, one enters the value for hypothesized difference,  $\delta_0$ ; 0 in this case. The resulting MINITAB outputs for this problem are displayed as follows:

### Two-Sample T-Test and CI: YA, YB

#### Two-sample T for YA vs YB

|    | N  | Mean  | StDev | SE Mean |
|----|----|-------|-------|---------|
| YA | 50 | 75.52 | 1.43  | 0.20    |
| YB | 50 | 72.47 | 2.76  | 0.39    |

Difference = mu (YA) - mu (YB)  
Estimate for difference: 3.047  
95% CI for difference: (2.169, 3.924)  
T-Test of difference = 0 (vs not =): T-Value = 6.92 P-Value = 0.000  
DF = 73

Several points are worth noting here:

1. The most important is the  $p$ -value which is virtually zero; the implication is that at the 0.05 significance level, we must reject the null hypothesis in favor of the alternative: the two population means are in fact different, i.e., the observed difference between the population is *not* zero. Note also that the  $t$ -statistic value is 6.92, a truly extreme value for a distribution that is symmetrical about the value 0, and for which the density value,  $f(t)$  essentially vanishes (i.e.,  $f(t) \approx 0$ ), for values of the  $t$  variate exceeding  $\pm 4$ . The  $p$ -value is obtained as  $P(|T| > 6.92)$ .
2. The estimated sample difference is 3.047, with a 95% confidence interval, (2.169, 3.924); since this interval does not contain the hypothesized difference  $\delta_0 = 0$ , the implication is that the test will reject  $H_0$ , as indeed we have concluded in point #1 above;
3. Finally, even though there were 50 data entries each for  $Y_A$  and  $Y_B$ , the degrees of freedom associated with this test is obtained as 73. (See the expressions in Eqs (15.59) and (15.60) above.)

This first test has therefore established that the means of the  $Y_A$  and  $Y_B$  populations are different, at the 5% significance level. Next, we wish to test which of these two different means is larger. To do this, the hypotheses to be tested are:

$$\begin{aligned} H_0 : \mu_A - \mu_B &= 0 \\ H_a : \mu_A &> \mu_B \end{aligned} \quad (15.62)$$

The resulting outputs from MINITAB are identical to what is shown above for the first test, except that the “95% CI for difference” line is replaced with 95% lower bound for difference: 2.313 and the “T-Test of difference = 0 (vs not =)” is replaced with T-Test of difference = 0 (vs >). The  $t$ -value,  $p$ -value and “DF” remain the same.

Again, with a  $p$ -value that is virtually zero, the conclusion is that, at the 5% significance level, the null hypothesis must be rejected in favor of the alternative, which, this time, is specifically that  $\mu_A$  is greater than  $\mu_B$ . Note that the value 2.313, computed from the data as the 95% lower bound for the difference, is considerably higher than the hypothesized value of 0; i.e., the hypothesized  $\delta_0 = 0$  lies well to the left of this lower bound for the difference. This is consistent with rejecting the null hypothesis in favor of the alternative, at the 5% significance level.

With the final test, we wish to sharpen the postulated difference a bit further. This time, we assert that,  $\mu_A$  is not only greater than  $\mu_B$ ; the former is in fact greater than the latter by a value that exceeds 2. The hypotheses are set up in this case as follows:

$$\begin{aligned} H_0 : \mu_A - \mu_B &= 2 \\ H_a : \mu_A &> \mu_B = 2 \end{aligned} \quad (15.63)$$

This time, in the MINTAB options, the new hypothesized difference is indicated as 2 in the “Test difference” box. The MINITAB results are displayed as follows:

### Two-Sample T-Test and CI: YA, YB

#### Two-sample T for YA vs YB

|    | N  | Mean  | StDev | SE Mean |
|----|----|-------|-------|---------|
| YA | 50 | 75.52 | 1.43  | 0.20    |
| YB | 50 | 72.47 | 2.76  | 0.39    |

```
Difference = mu (YA) - mu (YB)
Estimate for difference: 3.047
95% lower bound for difference: 2.313
T-Test of difference = 2 (vs >): T-Value = 2.38 P-Value = 0.010 DF
= 73
```

Note that the  $t$ -value is now 2.38 (reflecting the new hypothesized value of

$\delta_0 = 2$ ), with the immediate consequence that the  $p$ -value is now 0.01; not surprisingly, everything else remains the same as in the first test. Thus, at the 0.05 significance level, we reject the null hypothesis in favor of the alternative. Note also that the 95% lower bound for the difference is larger than the hypothesized difference of 2.

The conclusion is therefore that, with 95% confidence (or alternatively at a significance level of 0.05), the mean yield obtainable from the challenger process A is *at least* 2 points larger than that obtainable by the incumbent process B.

### 15.4.3 Paired Differences

A subtle but important variation on the theme of inference concerning two normal population means arises when the data naturally occur in pairs, as with the data shown in Table 15.6. This is a record of the “before” and “after” weights (in pounds) of twenty patients enrolled in a clinically-supervised 10-week weight-loss program. Several important characteristics set this problem apart from the general two-sample problem:

1. For each patient, the random variable “Weight” naturally occurs as an ordered pair of random variables  $(X, Y)$ , with  $X$  as the “Before” weight, and  $Y$  as the “After” weight;
2. As a result, it is highly unlikely that the two entries per patient will be totally independent, i.e., the random sample,  $X_1, X_2, \dots, X_n$ , will likely *not* be independent of  $Y_1, Y_2, \dots, Y_n$ ;
3. In addition, the sample sizes for each random sample,  $X_1, X_2, \dots, X_n$ , and  $Y_1, Y_2, \dots, Y_n$ , by definition, will be identical;
4. Finally, it is quite possible that the patient-to-patient variability in each random variable  $X$  or  $Y$  (i.e., the variability *within* each group) may be much larger than the difference *between* the groups that we seek to detect.

These circumstances call for a different approach, especially in light of item #2 above, which invalidates one of the most crucial assumptions underlying the two-sample tests: independence of the random samples.

The analysis for this class of problems proceeds as follows: Let  $(X_i, Y_i); i = 1, 2, \dots, n$ , be an ordered pair of random samples, where  $X_1, X_2, \dots, X_n$  is from a normal population with mean,  $\mu_X$ , and variance,  $\sigma_X^2$ ; and  $Y_1, Y_2, \dots, Y_n$ , a random sample from a normal population with mean,  $\mu_Y$ , and variance,  $\sigma_Y^2$ . Define the difference  $D$  as:

$$D_i = X_i - Y_i \quad (15.64)$$

then,  $D_i, i = 1, 2, \dots, n$ , constitutes a random sample of differences with mean value,

$$\delta = \mu_X - \mu_Y \quad (15.65)$$

**TABLE 15.6:** “Before” and “After” weights for patients on a supervised weight-loss program

| Patient #       | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  |
|-----------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Before Wt (lbs) | 272 | 319 | 253 | 325 | 236 | 233 | 300 | 260 | 268 | 276 |
| After Wt (lbs)  | 263 | 313 | 251 | 312 | 227 | 227 | 290 | 251 | 262 | 263 |
| Patient #       | 11  | 12  | 13  | 14  | 15  | 16  | 17  | 18  | 19  | 20  |
| Before Wt (lbs) | 215 | 245 | 248 | 364 | 301 | 203 | 197 | 217 | 210 | 223 |
| After Wt (lbs)  | 206 | 235 | 237 | 350 | 288 | 195 | 193 | 216 | 202 | 214 |

The quantities required for the hypothesis test are: the sample average,

$$\bar{D} = \frac{\sum_{i=1}^n D_i}{n}, \quad (15.66)$$

(which is unbiased for  $\delta$ ), and the sample variance of the differences,

$$S_D^2 = \frac{\sum_{i=1}^n (D_i - \bar{D})^2}{n-1}. \quad (15.67)$$

Under these circumstances, the null hypothesis is defined as

$$H_0 : \delta = \delta_0 \quad (15.68)$$

when  $\delta$ , the difference between the paired observations, is postulated as some value  $\delta_0$ . This hypothesis, as usual, is to be tested against the possible alternatives

$$\text{Lower-tailed } H_a : \delta < \delta_0 \quad (15.69)$$

$$\text{Upper-tailed } H_a : \delta > \delta_0 \quad (15.70)$$

$$\text{Two-tailed } H_a : \delta \neq \delta_0 \quad (15.71)$$

The appropriate test statistic is

$$T = \frac{\bar{D} - \delta_0}{S_D / \sqrt{n}}; \quad (15.72)$$

it possesses a  $t(n-1)$  distribution. When used to carry out what is generally known as the “paired  $t$ -test,” the results are similar to those obtained for earlier tests, with the specific rejection conditions summarized in Table 15.7. The next two examples illustrate the importance of distinguishing between a paired-test and a general two-sample test.

#### Example 15.7: WEIGHT-LOSS DATA ANALYSIS: PART 1

By treating the weight-loss patient data in Table 15.6 as “before” and “after” ordered pairs, determine at the 5% level, whether or not the weight loss program has been effective in assisting patients lose weight.

**TABLE 15.7:** Summary of  $H_0$  rejection conditions for the paired  $t$ -test

| Testing Against              | For general $\alpha$<br>Reject $H_0$ if:                                    |
|------------------------------|---|
| $H_a : \delta < \delta_0$    | $t < -t_\alpha(\nu)$  |
| $H_a : \delta > \delta_0$    | $t > t_\alpha(\nu)$   |
| $H_a : \delta \neq \delta_0$ | $t < -t_{\alpha/2}(\nu)$ or<br>$t > t_{\alpha/2}(\nu)$<br>( $\nu = n - 1$ ) |

**Solution:**

This problem requires determining whether the mean difference between the “before” and “after” weights for the 20 patients is significantly different from zero. The null and alternative hypotheses are:

$$\begin{aligned} H_0 : \delta &= 0 \\ H_a : \delta &\neq 0 \end{aligned} \quad (15.73)$$

We can compute the twenty “before”-minus-“after” weight differences, obtain the sample average and sample standard deviation of these differences, and then compute the  $t$ -statistic from Eq (15.72) for  $\delta_0 = 0$ . How this  $t$  statistic compares against the critical value of  $t_{0.025}(19)$  will determine whether or not to reject the null hypothesis.

We can also use MINITAB directly. After entering the data into two columns “Before WT” and “After WT”, the sequence: **Stat** > **Basic Statistics** > **Paired t** opens the usual analysis dialog box: as with other hypothesis tests, data columns are identified, and with the **OPTIONS** button, the “Alternative” for  $H_a$  is selected as “not equal,” along with 0 for the “Test mean” value, with the default confidence level (95.0). The resulting MINITAB outputs for this problem are displayed as follows:

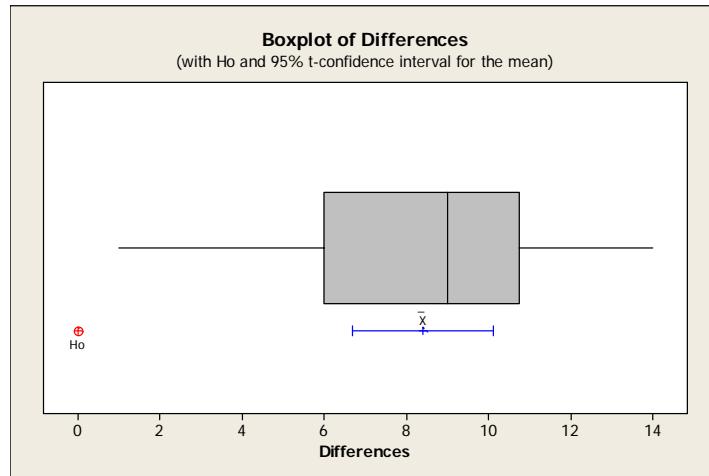
**Paired T-Test and CI: Before WT, After WT**

| Paired T for Before WT - After WT |    |       |       |         |
|-----------------------------------|----|-------|-------|---------|
|                                   | N  | Mean  | StDev | SE Mean |
| Before WT                         | 20 | 258.2 | 45.2  | 10.1    |
| After WT                          | 20 | 249.9 | 43.3  | 9.7     |
| Difference                        | 20 | 8.400 | 3.662 | 0.819   |

95% CI for mean difference: (6.686, 10.114)

T-Test of mean difference = 0 (vs not = 0): T-Value = 10.26 P-Value = 0.000

The mean difference (i.e., average weight-loss per patient) is 8.4 lbs, and the 95% confidence interval (6.686, 10.114), does not contain 0; also, the  $p$ -value is 0 (to three decimal places). The implication is therefore that at



**FIGURE 15.8:** Box plot of differences between the “Before” and “After” weights, including a 95% confidence interval for the mean difference, and the hypothesized  $H_0$  point,  $\delta_0 = 0$

the significance level of 0.05, we reject the null hypothesis and conclude that the weight-loss program was effective. The average weight loss of 8.4 lbs is therefore significantly different from zero, at the 5% significance level.

A box plot of the differences between the “before” and “after” weights is shown in Fig 15.8, which displays graphically that the null hypothesis should be rejected in favor of the alternative. Note how far the hypothesized value of 0 is from the 95% confidence interval for the mean weight difference.

The next example illustrates the consequences of wrongly employing a two-sample  $t$ -test for this natural paired  $t$ -test problem.

**Example 15.7: WEIGHT-LOSS DATA ANALYSIS: PART 2:  
TWO-SAMPLE T-TEST**

Revisit the problem in Example 15.6 but this time treat the “before” and “after” weight data in Table 15.6 as if they were independent samples from two different normal populations; carry out a 2-sample  $t$ -test and, at the 5% level, determine whether or not the two sample means are different.

**Solution:**

First let us be very clear: this is *not* the right thing to do; but if a 2-sample  $t$ -test is carried out on this data set with the hypotheses as:

$$\begin{aligned} H_0 : \mu_{\text{before}} - \mu_{\text{after}} &= 0 \\ H_a : \mu_{\text{before}} - \mu_{\text{after}} &\neq 0 \end{aligned} \quad (15.74)$$

MINITAB produces the following result:

**Two-Sample T-Test and CI: Before WT, After WT**

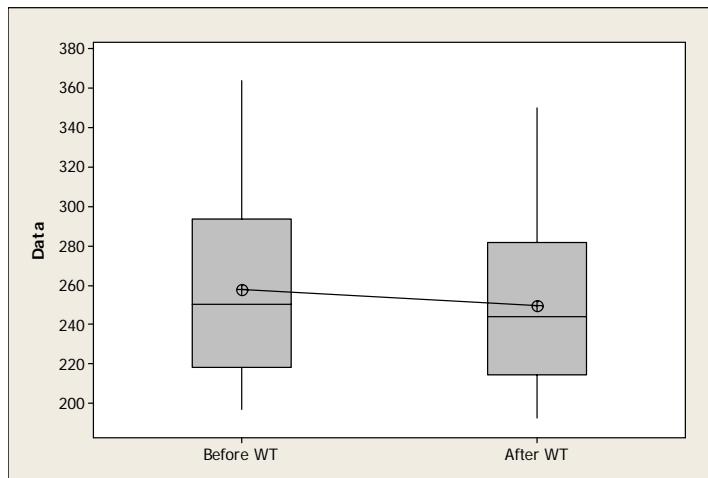
| <b>Two-sample T for Before WT vs After WT</b> |    |       |       |         |
|---|----|-------|-------|---------|
|   | N  | Mean  | StDev | SE Mean |
| Before WT                                     | 20 | 258.2 | 45.2  | 10.1    |
| After WT                                      | 20 | 249.9 | 43.3  | 9.7     |

```
Difference = mu (Before WT) - mu (After WT)
Estimate for difference: 8.4
95% CI for difference: (-20.0, 36.8)
T-Test of difference = 0 (vs not =): T-Value = 0.60 P-Value =
0.552 DF = 38
Both use Pooled StDev = 44.2957
```

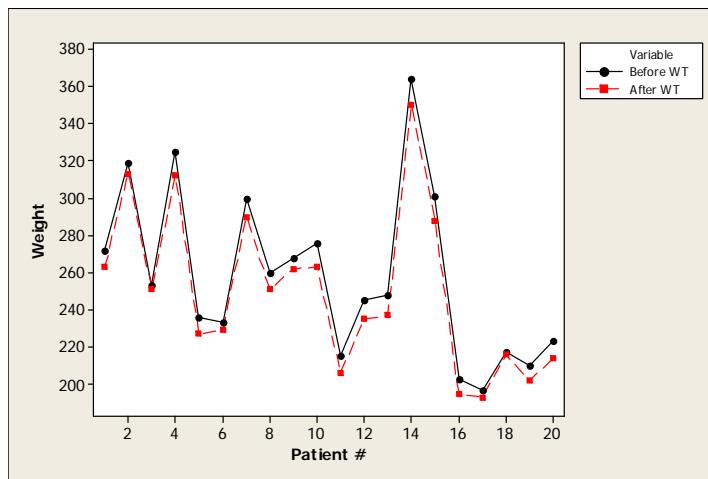
With a *t*-value of 0.6 and a *p*-value of 0.552, this analysis indicates that there is no evidence to support rejecting the null hypothesis at the significance level of 0.05. The estimated *difference* of the means is 8.4 (the same as the *mean* of the differences obtained in Example 15.6); but because of the large pooled standard deviation, the 95% confidence interval is (-20.0, 36.8), which includes 0. As a result, the null hypothesis cannot be rejected at the 5% significance level in favor of the alternative. This, of course, will be the wrong decision (as the previous example has shown) and should serve as a warning against using the two-sample *t*-test improperly for paired data.

It is important to understand the sources of the failure in this last example. First, a box plot of the two data sets, shown in Fig 15.9, graphically illustrates why the two-sample *t*-test is entirely unable to detect the very real, and very significant, difference between the “before” and “after” weights. The variability *within* the samples is so high that it swamps out the difference *between* each pair which is actually significant. But the most important reason is illustrated in Fig 15.10, which shows a plot of “before” and “after” weights for each patient versus patient number, from where it is absolutely clear, that the two sets of weights are almost perfectly correlated. Paired data are often *not* independent. Observe from the data (and from this graph) that without exception, every single “before” weight is higher than the corresponding “after” weight. The issue is therefore not whether there is a weight loss; it is a question of how much. For this group of patients, however, this difference cannot be detected in the midst of the large amount of variability *within* each group (“before” or “after”).

These are the primary reasons that the two-sample *t*-test failed miserably in identifying a differential that is quite significant. (As an exercise, the reader should obtain a scatter plot of the “before” weight versus the “after” weight to provide further graphical evidence of just how correlated the two weights are.)



**FIGURE 15.9:** Box plot of the “Before” and “After” weights including individual data means. Notice the wide range of each data set



**FIGURE 15.10:** A plot of the “Before” and “After” weights for each patient. Note how one data sequence is almost perfectly correlated with the other; in addition note the relatively large variability intrinsic in each data set compared to the difference between each point

## 15.5 Determining $\beta$ , Power, and Sample Size

Determining  $\beta$ , the Type II error risk, and hence  $(1 - \beta)$ , the power of any hypothesis test, depends on whether the test is one- or two-sided. The same is also true of the complementary problem: the determination of experimental sample sizes required to achieve a certain pre-specified power. We begin our discussion of such issues with the one-sided test, specifically the upper-tailed test, with the null hypothesis as in Eq (15.16) and the alternative in Eq (15.18). The results for the lower-tailed, and the two-sided tests, which follow similarly, will be given without detailed derivations.

### 15.5.1 $\beta$ and Power

To determine  $\beta$  (and hence power) for the upper-tailed test, it is not sufficient merely to state that  $\mu > \mu_0$ ; instead, one must specify a particular value for the alternative mean, say  $\mu_a$ , so that:

$$H_a : \mu = \mu_a > \mu_0 \quad (15.75)$$

is the alternative hypothesis. The Type II error risk is therefore the probability of failing to reject the null hypothesis when in truth the data came from the alternative distribution with mean  $\mu_a$  (where, for the upper-tailed test,  $\mu_a > \mu_0$ ).

The difference between this alternative and the postulated null hypothesis distribution mean,

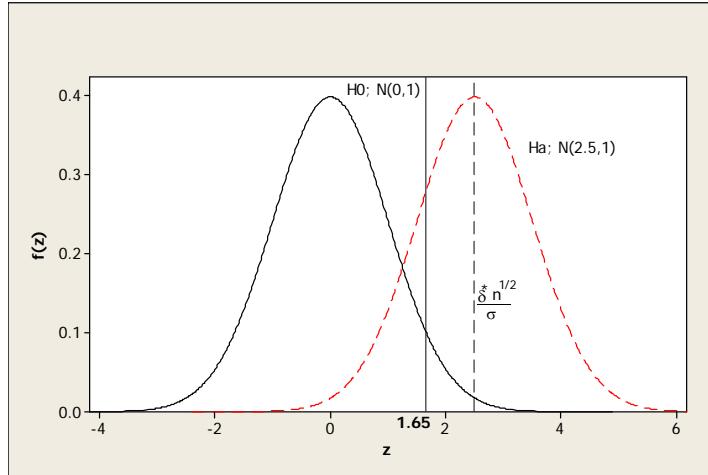
$$\delta^* = \mu_a - \mu_0 \quad (15.76)$$

is the margin by which the null hypothesis is falsified in comparison to the alternative. As one might expect, the magnitude of  $\delta^*$  will be a factor in how easy or difficult it is for the test to detect, amidst all the variability in the data, a difference between  $H_0$  and  $H_a$ , and therefore correctly reject  $H_0$  when it is false. (Equivalently, the magnitude of  $\delta^*$  will also factor into the risk of incorrectly failing to reject  $H_0$  in favor of a true  $H_a$ .)

As shown earlier, if  $H_0$  is true, then the distribution of the sample mean,  $\bar{X}$ , is  $N(\mu_0, \sigma^2/n)$ , so that the test statistic,  $Z$ , in Eq (15.20), possesses a standard normal distribution; i.e.,

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1) \quad (15.77)$$

However, if  $H_a$  is true, then in fact the more appropriate distribution for  $\bar{X}$  is  $N(\mu_a, \sigma^2/n)$ . And now, because  $E(\bar{X}) = \mu_a$  under these circumstances, not  $\mu_0$  as postulated, the most important implication is that the distributional characteristics of the computed  $Z$  statistic, instead of following the standard



**FIGURE 15.11:** Null and alternative hypotheses distributions for upper-tailed test based on  $n = 25$  observations, with population standard deviation  $\sigma = 4$ , where the true alternative mean,  $\mu_a$ , exceeds the hypothesized one by  $\delta^* = 2.0$ . The figure shows a “*z*-shift” of  $(\delta^* \sqrt{n})/\sigma = 2.5$ ; and with reference to  $H_0$ , the critical value  $z_{0.05} = 1.65$ . The area under the  $H_0$  curve to the *right* of the point  $z = 1.65$  is  $\alpha = 0.05$ , the significance level; the area under the dashed  $H_a$  curve to the *left* of the point  $z = 1.65$  is  $\beta$

normal distribution, will be:

$$Z \sim N \left( \frac{\delta^*}{\sigma/\sqrt{n}}, 1 \right) \quad (15.78)$$

i.e., the standard normal distribution shifted to the right (for this upper-tailed test) by a factor of  $(\delta^* \sqrt{n})/\sigma$ . Thus, as a result of a true differential,  $\delta^*$ , between alternative and null hypothesized means, the standardized alternative distribution will show a “*z*-shift”

$$z_{shift} = \frac{\delta^* \sqrt{n}}{\sigma} \quad (15.79)$$

For example, for a test based on 25 observations, with population standard deviation  $\sigma = 4$  where the true alternative mean,  $\mu_a$ , exceeds the hypothesized one by  $\delta^* = 2.0$ , the mean value of the standardized alternative distribution, following Eq (15.78), will be 2.5, and the two distributions will be as shown in Fig 15.11, with the alternative hypothesis distribution shown with the dashed line.

In terms of the standard normal variate,  $z$ , under  $H_0$ , the shifted variate under the alternative hypothesis,  $H_a$ , is:

$$\zeta = z - \frac{\delta^* \sqrt{n}}{\sigma} \quad (15.80)$$

And now, to compute  $\beta$ , we recall that, by definition,

$$\beta = P(z < z_\alpha | H_a) \quad (15.81)$$

which, by virtue of the “z-shift” translates to:

$$\beta = P\left(z < z_\alpha - \frac{\delta^* \sqrt{n}}{\sigma}\right), \quad (15.82)$$

from where we obtain the expression for the power of the test as:

$$(1 - \beta) = 1 - P\left(z < z_\alpha - \frac{\delta^* \sqrt{n}}{\sigma}\right) \quad (15.83)$$

Thus, for the illustrative example test given above, based on 25 observations, with  $\sigma = 4$  and  $\mu_a - \mu_0 = \delta^* = 2.0$ , the  $\beta$ -risk and power are obtained as

$$\begin{aligned} \beta &= P(z < 1.65 - 2.5) = 0.198 \\ \text{Power} &= (1 - \beta) = 0.802 \end{aligned} \quad (15.84)$$

as shown in Fig 15.12.

### 15.5.2 Sample Size

In the same way in which  $z_\alpha$  was defined earlier, let  $z_\beta$  be the standard normal variate such that

$$P(z > z_\beta) = \beta \quad (15.85)$$

so that, by symmetry,

$$P(z < -z_\beta) = \beta \quad (15.86)$$

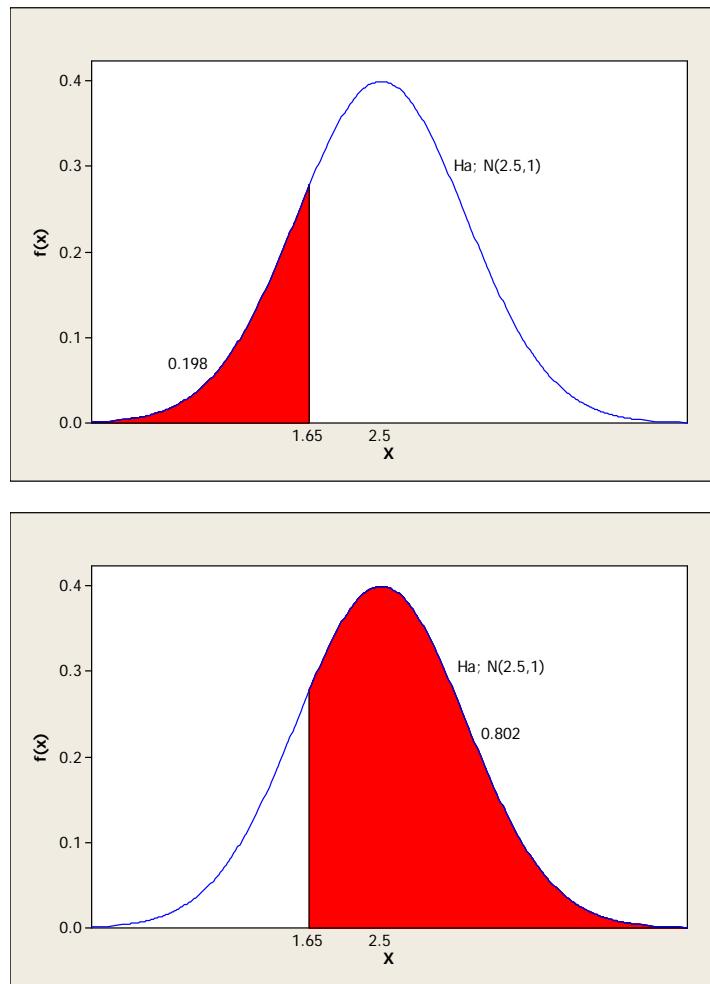
Then, from Eqs (15.82) and (15.86) we obtain :

$$-z_\beta = z_\alpha - \frac{\delta^* \sqrt{n}}{\sigma} \quad (15.87)$$

which rearranges to give the important expression,

$$z_\alpha + z_\beta = \frac{\delta^* \sqrt{n}}{\sigma} \quad (15.88)$$

which relates the  $\alpha$ - and  $\beta$ -risk variates to the three hypothesis test characteristics:  $\delta^*$ , the hypothesized mean shift to be detected by the test (the “signal”);  $\sigma$ , the population standard deviation, a measure of the variability inherent in the data (the “noise”); and finally,  $n$ , the number of experimental observations to be used to carry out the hypothesis test (the “sample size”). (Note that these three terms comprise what we earlier referred to as the “z-shift,” the precise amount by which the standardized  $H_a$  distribution has been shifted away from the  $H_0$  distribution: see Fig 15.11.)



**FIGURE 15.12:**  $\beta$  and power values for hypothesis test of Fig 15.11 with  $H_a \sim N(2.5, 1)$ . Top:  $\beta$ ; Bottom: Power =  $(1 - \beta)$

This relationship, fundamental to power and sample size analyses, can also be derived in terms of the unscaled critical value,  $x_C$ , which marks the boundary of the rejection region for the unscaled sample mean.

Observe that by definition of the significance level,  $\alpha$ , the critical value, and the  $Z$  statistic,

$$z_\alpha = \frac{x_C - \mu_0}{\sigma/\sqrt{n}} \quad (15.89)$$

so that:

$$x_C = z_\alpha \frac{\sigma}{\sqrt{n}} + \mu_0 \quad (15.90)$$

By definition of  $\beta$ , under  $H_a$ ,

$$\beta = P \left( z < \frac{x_C - \mu_a}{\sigma/\sqrt{n}} \right) \quad (15.91)$$

and from the definition of the  $z_\beta$  variate in Eq (15.86), we obtain:

$$-z_\beta = \frac{x_C - \mu_a}{\sigma/\sqrt{n}} \quad (15.92)$$

and upon substituting Eq (15.90) in for  $x_C$ , and recalling that  $\mu_a - \mu_0 = \delta^*$ , Eq (15.92) immediately reduces to

$$\begin{aligned} -z_\beta &= z_\alpha - \frac{\delta^* \sqrt{n}}{\sigma}, \text{ or} \\ z_\alpha + z_\beta &= \frac{\delta^* \sqrt{n}}{\sigma} \end{aligned} \quad (15.93)$$

as obtained earlier from the standardized distributions.

Several important characteristics of hypothesis tests are embedded in this important expression that are worth drawing out explicitly; but first, a general statement regarding  $z$ -variates and risks. Observe that any tail area,  $\tau$ , decreases as  $|z_\tau|$  increases; similarly, tail area,  $\tau$ , increases as  $z_\tau$  decreases; similarly, tail area,  $\tau$ , increases as  $|z_\tau|$  decreases. We may thus note the following about Eq (15.93):

1. The equation shows that for any particular hypothesis test with fixed characteristics  $\delta^*$ ,  $\sigma$ , and  $n$ , there is a conservation of the sum of the  $\alpha$ - and  $\beta$ -risk variates; if  $z_\alpha$  increases,  $z_\beta$  must decrease by a commensurate amount, and vice versa.
2. Consequently, if, in order to reduce the  $\alpha$ -risk,  $z_\alpha$  is increased,  $z_\beta$  will decrease commensurately to maintain the left-hand side sum constant, with the result that the  $\beta$ -risk must automatically increase. The reverse is also true: increasing  $z_\beta$  for the purpose of reducing the  $\beta$ -risk will result in  $z_\alpha$  decreasing to match the increase in  $z_\beta$ , so that the  $\alpha$  risk will then increase. Therefore, *for a fixed set of test characteristics, the associated Type I and Type II risks are such that a reduction in one risk will result in an increase in the other in mutual fashion.*

3. The only way to reduce either risk *simultaneously* (which will require increasing the total sum of the risk variates) is by increasing the “z-shift.” This is achievable most directly by increasing  $n$ , the sample size, since neither  $\sigma$ , the population standard deviation, nor  $\delta^*$ , the hypothesized mean shift to be detected by the test, is usually under the direct control of the experimenter.

This last point leads directly to the issue of determining how many experimental samples are required to attain a certain power, given basic test characteristics. This question is answered by solving Eq (15.88) explicitly for  $n$  to obtain:

$$n = \left[ \frac{(z_\alpha + z_\beta)\sigma}{\delta^*} \right]^2 \quad (15.94)$$

Thus, by specifying the desired  $\alpha$ - and  $\beta$ -risks along with the test characteristics,  $\delta^*$ , the hypothesized mean shift to be detected by the test, and  $\sigma$ , the population standard deviation, one can use Eq (15.94) to determine the sample size required to achieve the desired risk levels. In particular, it is customary to specify the risks as  $\alpha = 0.05$  and  $\beta = 0.10$ , in which case,  $z_\alpha = z_{0.05} = 1.645$ ; and  $z_\beta = z_{0.10} = 1.28$ . Eq (15.94) then reduces to:

$$n = \left( \frac{2.925\sigma}{\delta^*} \right)^2 \quad (15.95)$$

from which, given  $\delta^*$  and  $\sigma$ , one can determine  $n$ .

**Example 15.8: SAMPLE SIZE REQUIRED TO IMPROVE POWER OF HYPOTHESIS TEST**

The upper-tailed hypothesis test illustrated in Fig 15.11 was shown in Eq (15.84) to have a power of 0.802 (equivalent to a  $\beta$ -risk of 0.182). It is based on a sample size of  $n = 25$  observations, population standard deviation  $\sigma = 4$ , and where the true alternative mean  $\mu_a$  exceeds the hypothesized one by  $\delta^* = 2.0$ . Determine the sample size required to improve the power from 0.802 to the customary 0.9.

**Solution:**

Upon substituting  $\sigma = 4$ ;  $\delta^* = 2$  into Eq (15.95), we immediately obtain  $n = 34.2$ , which should be rounded up to the nearest integer to yield 35. This is the required sample size, an increase of 10 additional observations. To compute the actual power obtained with  $n = 35$  (since it is technically different from the precise, but impractical,  $n = 34.2$  obtained from Eq (15.95)), we introduce  $n = 35$  in Eq (15.94) and obtain the corresponding  $z_\beta$  as 1.308; from here we may obtain  $\beta$  from MINITAB's cumulative probability feature as  $\beta = 0.095$ , and hence

$$\text{Power} = 1 - \beta = 0.905 \quad (15.96)$$

is the actual power.

### Practical Considerations

In practice, prior to performing the actual hypothesis test, no one knows whether or not  $H_a$  is true compared to  $H_0$  — talk less of knowing the precise amount by which  $\mu_a$  will exceed the postulated  $\mu_0$  if  $H_a$  turns out to be true. The implication therefore is that  $\delta^*$  is never known in an objective fashion *à-priori*. In determining the power of a hypothesis test, therefore,  $\delta^*$  is treated not as “known” but as a *design parameter*: the minimum difference we would like to detect, if such a difference exists. Thus,  $\delta^*$  is to be considered properly as the magnitude of the smallest difference we wish to detect with the hypothesis test.

In a somewhat related vein, the population standard deviation,  $\sigma$ , is rarely known *à priori* in many practical cases. Under these circumstances, it has often been recommended to use educated guesses, or results from prior experiments under similar circumstances, to provide pragmatic surrogates for  $\sigma$ . We strongly recommend an alternative approach: casting the problem in terms of the “signal-to-noise” ratio (SNR):

$$\rho_{SN} = \frac{\delta^*}{\sigma} \quad (15.97)$$

a ratio of the magnitude of the “signal” (difference in the means) to be detected and the intrinsic “noise” (population standard deviation) in the midst of which the signal is to be detected. In this case, the general equation (15.94), and the more specific Eq (15.95) become:

$$\begin{aligned} n &= \left[ \frac{(z_\alpha + z_\beta)}{\rho_{SN}} \right]^2 \\ n &= \left( \frac{2.925}{\rho_{SN}} \right)^2 \end{aligned} \quad (15.98)$$

Without necessarily knowing either  $\delta^*$  or  $\sigma$  independently, the experimenter then makes a sample-size decision by designing for a test to handle a “design” SNR.

**Example 15.9: SAMPLE SIZE TABLE FOR VARIOUS SIGNAL-TO-NOISE RATIOS: POWER OF 0.9**

Obtain a table of sample sizes required to achieve a power of 0.9 for various signal-to-noise ratios from 0.3 to 1.5.

**Solution:**

Table 15.8 is generated from Eq (15.98) for the indicated values of the signal-to-noise ratio, where  $n^+$  is the value of the computed  $n$  rounded up to the nearest integer. As expected, as the signal-to-noise ratio improves, the sample size required to achieve a power of 0.9 reduces; fewer data points are required to detect signals that are large relative to the standard deviation. Note in particular that for the example considered

**TABLE 15.8:** Sample size  $n$  required to achieve a power of 0.9 for various values of signal-to-noise ratio,  $\rho_{SN}$ 

| $\rho_{SN}$ | 0.3   | 0.4   | 0.5   | 0.6   | 0.7   | 0.8   | 0.9   | 1.0  | 1.2  | 1.5  |
|-------------|-------|-------|-------|-------|-------|-------|-------|------|------|------|
| $n$         | 95.06 | 53.47 | 34.22 | 23.77 | 17.46 | 13.37 | 10.56 | 8.56 | 5.94 | 3.80 |
| $n^+$       | 96    | 53    | 35    | 24    | 18    | 14    | 11    | 9    | 6    | 4    |

in Fig 15.11 and Example 15.8,  $\rho_{SN} = 2/4 = 0.5$ ; from this Table 15.8, the required sample size, 35, is precisely as obtained in Example 15.8.

### 15.5.3 $\beta$ and Power for Lower-Tailed and Two-Sided Tests

For the sake of clarity, the preceding discussion was specifically restricted to the upper-tailed test. Now that we have presented and illustrated the essential concepts, it is relatively straightforward to extend them to other types of tests without having to repeat the details.

First, because the sampling distribution for the test statistic employed for these hypothesis tests is symmetric, it is easy to see that with the lower-tailed alternative

$$H_a : \mu = \mu_a < \mu_0 \quad (15.99)$$

this time, with

$$\delta^* = \mu_0 - \mu_a, \quad (15.100)$$

the  $\beta$  risk is obtained as:

$$\beta = P\left(z > z_\alpha + \frac{\delta^* \sqrt{n}}{\sigma}\right) \quad (15.101)$$

the equivalent of Eq (15.82), from where the power is obtained as  $(1 - \beta)$ . Again, because of symmetry, it is easy to see that the expression for determining sample size is precisely the same as derived earlier for the upper tailed test: i.e.,

$$n = \left[ \frac{(z_\alpha + z_\beta)\sigma}{\delta^*} \right]^2$$

All other results therefore follow.

For the two-tailed test, things are somewhat different, of course, but the same principles apply. The  $\beta$  risk is determined from:

$$\beta = P\left(z < z_{\alpha/2} - \frac{\delta^* \sqrt{n}}{\sigma}\right) - P\left(z < -z_{\alpha/2} - \frac{\delta^* \sqrt{n}}{\sigma}\right) \quad (15.102)$$

because of the two-sided rejection region. Unfortunately, as a result of the additional term in this equation, there is no closed-form solution for  $n$  that is

the equivalent of Eq (15.94). When  $P\left(z < -z_{\alpha/2} - \frac{\delta^* \sqrt{n}}{\sigma}\right) \ll \beta$ , the approximation,

$$n \approx \left[ \frac{(z_{\alpha/2} + z_{\beta})\sigma}{\delta^*} \right]^2 \quad (15.103)$$

is usually good enough. Of course, given the test characteristics, computer programs can solve for  $n$  precisely in Eq (15.102) without the need to resort to the approximation shown here.

#### 15.5.4 General Power and Sample Size Considerations

For general power and sample size considerations, it is typical to start by specifying  $\alpha$  and  $\sigma$ ; as a result, in either Eq (15.94) for one-tailed tests, or Eq (15.103) for the two-sided test, this leaves 3 parameters to be determined:  $\delta^*$ ,  $n$ , and  $z_{\beta}$ . By specifying any two, a value for the third unspecified parameter that is consistent with the given information can be computed from these equations.

In MINITAB the sequence required for carrying out this procedure is: **Stat** > **Power and Sample Size** which produces a drop down menu containing a collection of hypothesis tests (and experimental designs — see later). Upon selecting the hypothesis test of interest, a dialog box opens, with the instruction to “Specify values for any two of the following,” with three appropriately labeled spaces for “Sample size(s),” “Difference(s),” and “Power value(s).” The “Options” button is used to specify the alternative hypothesis and the  $\alpha$ -risk value. The value of the unspecified third parameter is then computed by MINITAB.

The following example illustrates this procedure.

**Example 15.10: POWER AND SAMPLE SIZE DETERMINATION USING MINITAB**

Use MINITAB to compute power and sample size for an upper-tailed, one sample  $z$ -test, with  $\sigma = 4$ , designed to detect a difference of 2, at the significance level of  $\alpha = 0.05$ : (1) if  $n = 25$ , determine the resulting power; (2) when the power is desired to be 0.9, determine the required sample size. (3) With a sample size of  $n = 35$ , determine the minimum difference that can be detected with a power of 0.9.

**Solution:**

- (1) Upon entering the given parameters into the appropriate boxes in the MINITAB dialog box, and upon choosing the appropriate alternative hypothesis, the MINITAB result is shown below:

**Power and Sample Size**  
1-Sample Z Test

```
Testing mean = null (versus > null)
Calculating power for mean = null + difference
```

|  |
|--|
| <u>Alpha = 0.05 Assumed standard deviation = 4</u> |
| Sample   |
| Difference      Size      Power                    |
| 2                25        0.803765                |

This computed power value is what we had obtained earlier.

(2) When the power is specified and the sample size removed, the MINITAB result is:

**Power and Sample Size**  
1-Sample Z Test

|  |
|--|
| Testing mean = null (versus > null)                |
| Calculating power for mean = null + difference     |
| <u>Alpha = 0.05 Assumed standard deviation = 4</u> |
| Sample      Target                                 |
| Difference      Size      Power      Actual Power  |
| 2                35        0.9        0.905440     |

This is exactly the same sample size value and the same actual power value we had obtained earlier.

(3) With  $n$  specified as 35 and the difference unspecified, the MINITAB result is:

**Power and Sample Size**  
1-Sample Z Test  
Testing mean = null (versus > null)  
Calculating power for mean = null + difference  
Alpha = 0.05 Assumed standard deviation = 4

|   |
|---|
| Sample      Target                              |
| Difference      Size      Power      Difference |
| 2                35        0.9        1.97861   |

The implication is that any difference greater than 1.98 can be detected at the desired power. A difference of 2.0 is therefore detectable at a power that is at least 0.9.

These results are all consistent with what we had obtained earlier.

## 15.6 Concerning Variances of Normal Populations

The discussions up until now have focused exclusively on hypothesis tests concerning the means of normal populations. But if we recall, for example, the earlier statements made regarding, say, the yield of process A, that  $Y_A \sim N(75.5, 1.5^2)$ , we see that in this statement is a companion assertion about the

**TABLE 15.9:** Summary of  $H_0$  rejection conditions for the  $\chi^2$ -test

| Testing Against                  | For general $\alpha$<br>Reject $H_0$ if:                                 |
|----------------------------------|--|
| $H_a : \sigma^2 < \sigma_0^2$    | $c^2 < \chi_{1-\alpha}^2(n-1)$   |
| $H_a : \sigma^2 > \sigma_0^2$    | $c^2 > \chi_\alpha^2(n-1)$   |
| $H_a : \sigma^2 \neq \sigma_0^2$ | $c^2 < \chi_{1-\alpha/2}^2(n-1)$<br>or<br>$c^2 > \chi_{\alpha/2}^2(n-1)$ |

associated variance. To confirm or refute this statement *completely* requires testing the validity of the assertion about the variance also.

There are two classes of tests concerning variances of normal population: the first concerns testing the variance obtained from a sample against a postulated population variance (as is the case here with  $Y_A$ ); the second concerns testing two (independent) normal populations for equality of their variances. We shall now deal with each case.

### 15.6.1 Single Variance

When the variance of a sample is to be tested against a postulated value,  $\sigma_0^2$ , the null hypothesis is:

$$H_0 : \sigma^2 = \sigma_0^2 \quad (15.104)$$

Under the assumption that the sample in question came from a normal population, then the test statistic:

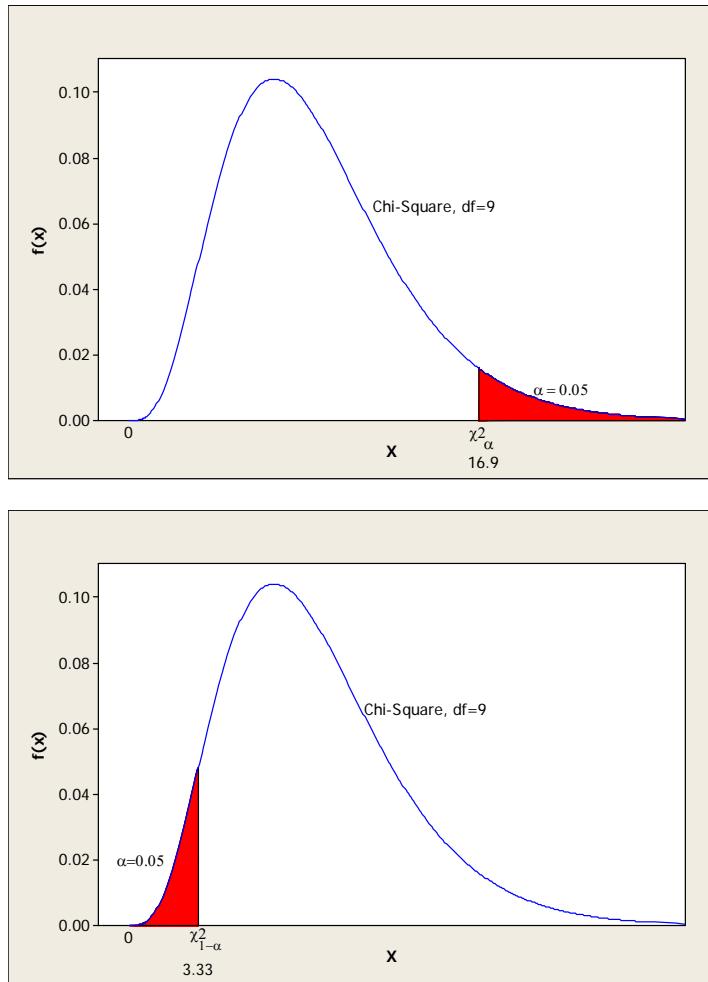
$$C^2 = \frac{(n-1)S^2}{\sigma_0^2} \quad (15.105)$$

has a  $\chi^2(n-1)$  distribution, if  $H_0$  is true. As a result, this test is known as a “chi-squared” test; and the rejection criteria for the usual triplet of alternatives is shown in Table 15.9. The reader should note the lack of symmetry in the boundaries of these rejection regions when compared with the symmetric boundaries for the corresponding  $z$ - and  $t$ -tests. This, of course, is a consequence of the asymmetry of the  $\chi^2(n-1)$  distribution. For example, for one-sided tests based on 10 samples from a normal distribution, the null hypothesis distributions for  $C^2$  is shown in Fig 15.13.

The next example is used to illustrate a two-sided test.

#### Example 15.11: VARIANCE OF “PROCESS A” YIELD

Formulate and test an appropriate hypothesis, at the significance level of 0.05, regarding the variance of the yield obtainable from process A implied by the assertion that the sample data presented in Chapter 1



**FIGURE 15.13:** Rejection regions for one-sided tests of a single variance of a normal population, at a significance level of  $\alpha = 0.05$ , based on  $n = 10$  samples. The distribution is  $\chi^2(9)$ ; Top: for  $H_a : \sigma^2 > \sigma_0^2$ , indicating rejection of  $H_0$  if  $c^2 > \chi^2_\alpha(9) = 16.9$ ; Bottom: for  $H_a : \sigma^2 < \sigma_0^2$ , indicating rejection of  $H_0$  if  $c^2 < \chi^2_{1-\alpha}(9) = 3.33$

for  $Y_A$  is from a normal population with the distribution  $N(75.5, 1.5^2)$ .

**Solution:**

The hypothesis to be tested is that  $\sigma_A^2 = 1.5^2$  against the alternative that it is not; i.e.,:

$$\begin{aligned} H_0 : \quad \sigma_A^2 &= 1.5^2 \\ H_a : \quad \sigma_A^2 &\neq 1.5^2, \end{aligned} \quad (15.106)$$

The sample variance computed from the supplied data is  $s_A^2 = 2.05$ , so that the specific value for the  $\chi^2$  test statistic is:

$$c^2 = \frac{49 \times 2.05}{2.25} = 44.63 \quad (15.107)$$

The rejection region for this two-sided test, with  $\alpha = 0.05$ , is shown in Fig 15.14, for a  $\chi^2(49)$  distribution. The boundaries of the rejection region are obtained from the usual cumulative probabilities; the left boundary is obtained by finding  $\chi_{1-\alpha/2}^2$  such that

$$\begin{aligned} P(c^2 > \chi_{1-\alpha/2}^2(49)) &= 0.975 \\ \text{or } P(c^2 < \chi_{1-\alpha/2}^2(49)) &= 0.025 \\ \text{i.e., } \chi_{1-\alpha/2}^2 &= 31.6 \end{aligned} \quad (15.108)$$

and the right boundary from:

$$\begin{aligned} P(c^2 > \chi_{\alpha/2}^2(49)) &= 0.025 \\ \text{or } P(c^2 < \chi_{\alpha/2}^2(49)) &= 0.975 \\ \text{i.e., } \chi_{\alpha/2}^2 &= 70.2 \end{aligned} \quad (15.109)$$

Since the value for  $c^2$  above does not fall into this rejection region, we do not reject the null hypothesis.

As before, MINITAB could be used directly to carry out this test. The self-explanatory procedure follows along the same lines as those discussed extensively above.

The conclusion: at the 5% significance level, we cannot reject the null hypothesis concerning  $\sigma_A^2$ .

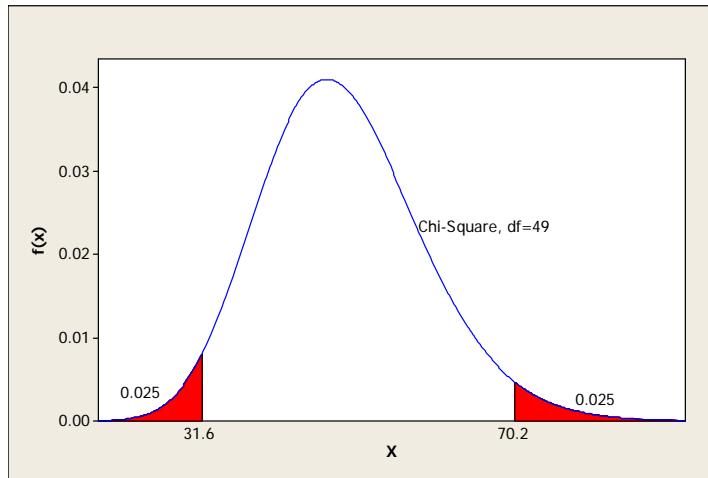
### 15.6.2 Two Variances

When two variances from mutually independent normal populations are to be compared, the null hypothesis is:

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad (15.110)$$

If the samples (sizes  $n_1$  and  $n_2$  respectively) come from independent normal distributions, then the test statistic:

$$F = \frac{S_1^2}{S_2^2} \quad (15.111)$$



**FIGURE 15.14:** Rejection regions for the two-sided tests concerning the variance of the process A yield data  $H_0 : \sigma_A^2 = 1.5^2$ , based on  $n = 50$  samples, at a significance level of  $\alpha = 0.05$ . The distribution is  $\chi^2(49)$ , with the rejection region shaded; because the test statistic,  $c^2 = 44.63$ , falls outside of the rejection region, we do not reject  $H_0$ .

has an  $F(\nu_1, \nu_2)$  distribution, where  $\nu_1 = (n_1 - 1)$  and  $\nu_2 = (n_2 - 1)$ , if  $H_0$  is true. Such tests are therefore known as “ $F$ -tests.” As with other tests, the rejection regions are determined from the  $F$ -distribution with appropriate degrees-of-freedom pairs on the basis of the desired significance level,  $\alpha$ . These are shown in Table 15.10.

It is often helpful in carrying out  $F$ -tests to recall the following property of the  $F$ -distribution:

$$F_{1-\alpha}(\nu_1, \nu_2) = \frac{1}{F_\alpha(\nu_2, \nu_1)} \quad (15.112)$$

an easy enough relationship to prove directly from the definition of the  $F$ -

**TABLE 15.10:** Summary of  $H_0$  rejection conditions for the  $F$ -test

| Testing Against                    | For general $\alpha$   |
|------------------------------------|--|
| $H_a : \sigma_1^2 < \sigma_2^2$    | Reject $H_0$ if:<br>$f < F_{1-\alpha}(\nu_1, \nu_2)$                         |
| $H_a : \sigma_1^2 > \sigma_2^2$    | $f > F_\alpha(\nu_1, \nu_2)$   |
| $H_a : \sigma_1^2 \neq \sigma_2^2$ | $f < F_{1-\alpha/2}(\nu_1, \nu_2)$<br>or<br>$f > F_{\alpha/2}(\nu_1, \nu_2)$ |

statistic in Eq (15.111). This relationship makes it possible to reduce the number of entries in old-fashioned  $F$ -tables. As we have repeatedly advocated in this chapter, however, it is most advisable to use computer programs for carrying out such tests.

**Example 15.12: COMPARING VARIANCES OF YIELDS  
FROM PROCESSES A AND B**

From the data supplied in Chapter 1 on the yields obtained from the two chemical processes A and B, test a hypothesis on the potential equality of these variances, at the 5% significance level.

**Solution:**

The hypothesis to be tested is that  $\sigma_A^2 = \sigma_B^2$  against the alternative that it is not; i.e.,:

$$\begin{aligned} H_0 : \sigma_A^2 &= \sigma_B^2 \\ H_a : \sigma_A^2 &\neq \sigma_B^2 \end{aligned} \quad (15.113)$$

From the supplied data, we obtain  $s_A^2 = 2.05$ , and  $s_B^2 = 7.62$ , so that the specific value for the  $F$ -test statistic is obtained as:

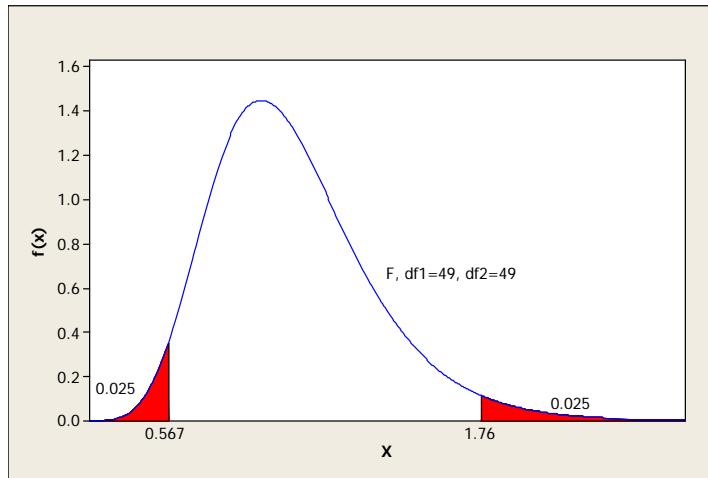
$$f = \frac{2.05}{7.62} = 0.27 \quad (15.114)$$

The rejection region for this two-sided  $F$ -test, with  $\alpha = 0.05$ , is shown in Fig 15.15, for an  $F(49, 49)$  distribution, with boundaries at  $f = 0.567$  to the left and 1.76 to the right, obtained as usual from cumulative probabilities. (Note that the value of  $f$  at one boundary is the reciprocal of the value at the other boundary.) Since the specific test value, 0.27, falls in the left side of the rejection region, we must therefore reject the null hypothesis in favor of the alternative that these two variances are unequal.

The self-explanatory procedure for carrying out the test in MINITAB generates results that include a  $p$ -value of 0.000, in agreement with the conclusion above to reject the null hypothesis at the 5% significance level.

The  $F$ -test is particularly useful for ascertaining whether or not the assumption of equality of variances is valid *before* performing a two-sample  $t$ -test. If the null hypothesis regarding the equality assumption is rejected, then one must not use the “equal variance” option of the test. If one is unable to reject the null hypothesis, one may proceed to use the “equal variance” option. As discussed in subsequent chapters, the  $F$ -test is also at the heart of ANOVA (ANalysis Of VAriance), a methodology that is central to much of statistical design of experiments and the systematic analysis of the resulting data statistical tests involving several means, and even regression analysis.

Finally, we note that the  $F$ -test is quite sensitive to the normality assumption: if this assumption is invalid, the test results will be unreliable. Note that the assumption of normality is not about the mean of the data but about



**FIGURE 15.15:** Rejection regions for the two-sided tests of the equality of the variances of the process A and process B yield data, i.e.,  $H_0 : \sigma_A^2 = \sigma_B^2$ , at a significance level of  $\alpha = 0.05$ , based on  $n = 50$  samples each. The distribution is  $F(49, 49)$ , with the rejection region shaded; since the test statistic,  $f = 0.27$ , falls within the rejection region to the left, we reject  $H_0$  in favor of  $H_a$ .

the raw data set itself. One must therefore be careful to ensure that this normality assumption is reasonable before carrying out an  $F$ -test. If the data is from non-normal distributions, most computer programs provide alternatives (based on non-parametric methods).

## 15.7 Concerning Proportions

As noted at the beginning of this chapter, a statistical hypothesis, in the most fundamental sense, is an assertion or statement about one or more populations; and the hypothesis test provides an objective means of ascertaining the truth or falsity of such a statement. So far, our discussions have centered essentially around normal populations because a vast majority of practical problems are of this form, or can be safely approximated as such. However, not all problems of practical importance involve sampling from normal populations; and the next section will broach this topic from a more general perspective. For now, we want to consider first a particular important class of problems involving sampling from a non-Gaussian population: hypotheses concerning proportions.

The general theoretical characteristics of problems of this kind were stud-

ied extensively in Chapter 8. Out of a total number of  $n$  samples examined for a particular attribute,  $X$  is the total number of (discrete) observations sharing the attribute in question;  $X/n$  is therefore the observed sample proportion sharing the attribute. Theoretically, the random variable,  $X$ , is known to follow the binomial distribution, characterized by the parameter  $p$ , the theoretical population proportion sharing the attribute (also known as the “probability of success”). Statements about such proportions are therefore statistical hypotheses concerning samples from binomial populations. Market/opinion surveys (such as the example used to open Chapter 14) where the proportion preferring a certain brand is of interest, and manufacturing processes where the concern is the proportion of defective products, provide the prototypical examples of problems of this nature. Hypotheses about the probability of successful embryo implantation in in-vitro fertilization (discussed in Chapter 7), or any other such binomial process probability, also fall into this category.

We deal first with hypotheses concerning single population proportions, and then hypotheses concerning two proportions. The underlying principles remain the same as with other tests: find the appropriate test statistic and its sampling distribution, and, given a specific significance level, use these to make probabilistic statements that will allow the determination of the appropriate rejection region.

### 15.7.1 Single Population Proportion

The problem of interest involves testing a hypothesis concerning a single binomial population proportion,  $p$ , given a sample of  $n$  items from which one observes  $X$  “successes” (the same as the detection of the attribute in question); the null hypothesis is:

$$H_0 : p = p_0 \quad (15.115)$$

with  $p_0$  as the specific value postulated for the population proportion. The usual three possible alternative hypotheses are:

$$H_a : p < p_0 \quad (15.116)$$

$$H_a : p > p_0 \quad (15.117)$$

and

$$H_a : p \neq p_0. \quad (15.118)$$

To determine an appropriate test statistic and its sampling distribution, we need to recall several characteristics of the binomial random variable from Chapter 8. First, the estimator,  $\Pi$ , defined as:

$$\Pi = \frac{X}{n}, \quad (15.119)$$

the mean number of successes, is unbiased for the binomial population parameter; the mean of the sampling distribution for  $\Pi$  is therefore  $p$ . Next, the

**TABLE 15.11:** Summary of  $H_0$  rejection conditions for the single-proportion  $z$ -test

| Testing Against    | For general $\alpha$<br>Reject $H_0$ if:        | For $\alpha = 0.05$<br>Reject $H_0$ if: |
|--------------------|---|---|
| $H_a : p < p_0$    | $z < -z_\alpha$                                 | $z < -1.65$                             |
| $H_a : p > p_0$    | $z > z_\alpha$                                  | $z < 1.65$                              |
| $H_a : p \neq p_0$ | $z < -z_{\alpha/2}$<br>or<br>$z > z_{\alpha/2}$ | $z < -1.96$<br>or<br>$z > 1.96$         |

variance of  $\Pi$  is  $\sigma_X^2/n^2$ , where

$$\sigma_X^2 = npq = np(1-p) \quad (15.120)$$

is the variance of the binomial random variable,  $X$ . Hence,

$$\sigma_\Pi^2 = \frac{p(1-p)}{n} \quad (15.121)$$

### Large Sample Approximations

From the Central Limit Theorem we know that, in the limit as  $n \rightarrow \infty$ , the sampling distribution of the mean of any population (including the binomial) tends to the normal distribution. The implication is that the statistic,  $Z$ , defined as:

$$Z = \frac{\frac{\Pi}{n} - p}{\sqrt{p(1-p)/n}} \quad (15.122)$$

has an approximate standard normal,  $N(0, 1)$ , distribution for large  $n$ .

The test statistic for carrying out the hypothesis test in Eq (15.115) versus any of the three alternatives is therefore:

$$Z = \frac{\Pi - p_0}{\sqrt{p_0(1-p_0)/n}} \quad (15.123)$$

a test statistic with precisely the same properties as those used for the standard  $z$ -test. The rejection conditions are identical to those shown in Table 15.2, which, when modified appropriately for the one-proportion test, is as shown in Table 15.11.

Since this test is predicated upon the sample being “sufficiently large,” it is important to ensure that this is indeed the case. A generally agreed upon objective criterion for ascertaining the validity of this approximation is that the interval

$$I_0 = p_0 \pm 3\sqrt{[p_0(1-p_0)]/n} \quad (15.124)$$

does not include 0 or 1. The next example illustrates these concepts.

**Example 15.13: EXAM TYPE PREFERENCE OF UNDERGRADUATE CHEMICAL ENGINEERING STUDENTS**

In the opening sections of Chapter 14, we reported the result of an opinion poll of 100 undergraduate chemical engineering students in the United States: 75 of the students prefer “closed-book” exams to “opened-book” ones. At the 5% significance level, test the hypothesis that the true proportion preferring “closed-book” exams is in fact 0.8, against the alternative that it is not.

**Solution:**

If the sample size is confirmed to be large enough, then this is a single proportion test which employs the  $z$ -statistic. The interval  $p_0 \pm 3\sqrt{[p_0(1 - p_0)]/n}$  in this case is  $0.8 \pm 0.12$ , or  $(0.68, 0.92)$ , which does not include 0 or 1; the sample size is therefore considered to be sufficiently large.

The hypothesis to be tested is therefore the two-sided

$$\begin{aligned} H_0 : p &= 0.8 \\ H_a : p &\neq 0.8; \end{aligned} \quad (15.125)$$

the  $z$ -statistic in this case is:

$$z = \frac{0.75 - 0.8}{\sqrt{(0.8 \times 0.2)/100}} = -1.25 \quad (15.126)$$

Since this value does not lie in the two-sided rejection region for  $\alpha = 0.05$ , we do not reject the null hypothesis.

MINITAB could be used to tackle this example problem directly. The self-explanatory sequence (when one chooses the "use test and interval based on normal distribution" option) produces the following result:

| Test and CI for One Proportion    |    |     |          |                      |         |         |
|-----------------------------------|----|-----|----------|----------------------|---------|---------|
| Test of $p = 0.8$ vs $p \neq 0.8$ |    |     |          |                      |         |         |
| Sample                            | X  | N   | Sample p | 95% CI               | Z-Value | P-Value |
| 1                                 | 75 | 100 | 0.750000 | (0.665131, 0.834869) | -1.25   | 0.211   |

Using the normal approximation.

As with similar tests discussed earlier, we see here that the 95% confidence interval for the parameter,  $p$ , contains the postulated  $p_0 = 0.8$ ; the associated  $p$ -value for the test (an unfortunate and unavoidable notational clumsiness that we trust will not confuse the reader unduly<sup>1</sup>) is 0.211, so that we do not reject  $H_0$  at the 5% significance level.

---

<sup>1</sup>The latter  $p$  of the “ $p$ -value” should not be confused with the binomial “probability of success” parameter

### Exact Tests

Even though it is customary to invoke the normal approximation in dealing with tests for single proportions, this is in fact not necessary. The reason is quite simple: if  $X \sim Bi(n, p)$ , then  $\Pi = X/n$  has a  $Bi(n, p/n)$  distribution. This fact can be used to compute the probability that  $\Pi = p_0$ , or any other value — providing the means for determining the boundaries of the various rejection regions, (given desired tail area probabilities), just as with the standard normal distribution, or any other standardized test distribution. Computer programs such as MINITAB provide options for obtaining exact  $p$ -values for the single proportion test that are based on exact binomial distributions.

When MINITAB is used to carry out the test in Example 15.13 above, this time without invoking the normal approximation option, the result is as follows:

#### Test and CI for One Proportion

Test of  $p = 0.8$  vs  $p \neq 0.8$

| Sample | X  | N   | Sample p | Exact                |         |
|--------|----|-----|----------|----------------------|---------|
|        |    |     |          | 95% CI               | P-Value |
| 1      | 75 | 100 | 0.750000 | (0.653448, 0.831220) | 0.260   |

The 95% confidence interval, which is now based on a binomial distribution, not a normal approximation, is now slightly different; the  $p$ -value, is also now slightly different, but the conclusion remains the same.

### 15.7.2 Two Population Proportions

In comparing two population proportions,  $p_1$  and  $p_2$ , as with the 2-sample tests of means from normal populations, the null hypothesis is:

$$H_0 : \Pi_1 - \Pi_2 = \delta_0 \quad (15.127)$$

where  $\Pi_1 = X_1/n_1$  and  $\Pi_2 = X_2/n_2$  are, respectively, the random proportions of “successes” obtained from population 1 and population 2, based on samples of respective sizes  $n_1$  and  $n_2$ . For example,  $\Pi_1$  could be the fraction of defective chips in a sample of  $n_1$  chips manufactured at one facility whose true proportion of defectives is  $p_1$ , while  $\Pi_2$  is the defective fraction contained in a sample from a different facility. The difference between the two population proportions is postulated as some value  $\delta_0$  that need not be zero.

As usual, the hypothesis is to be tested against the possible alternatives:

$$\text{Lower-tailed } H_a : \Pi_1 - \Pi_2 < \delta_0 \quad (15.128)$$

$$\text{Upper-tailed } H_a : \Pi_1 - \Pi_2 > \delta_0 \quad (15.129)$$

$$\text{Two-tailed } H_a : \Pi_1 - \Pi_2 \neq \delta_0 \quad (15.130)$$

As before,  $\delta_0 = 0$  constitutes a test of equality of the two proportions.

To obtain an appropriate test statistic and its sampling distribution, we begin by defining:

$$D_{\Pi} = \Pi_1 - \Pi_2 \quad (15.131)$$

We know in general that

$$E(D_{\Pi}) = \mu_{D_{\Pi}} = p_1 - p_2 \quad (15.132)$$

$$\sigma_{D_{\Pi}} = \sqrt{\left(\frac{p_1 q_1}{n_1}\right) + \left(\frac{p_2 q_2}{n_2}\right)} \quad (15.133)$$

But now, if the sample sizes  $n_1$  and  $n_2$  are large, then it can be shown that

$$D_{\Pi} \sim N(\mu_{D_{\Pi}}, \sigma_{D_{\Pi}}^2) \quad (15.134)$$

again allowing us to invoke the normal approximation (for large sample sizes). This immediately implies that the following is an appropriate test statistic to use for this two-proportion test:

$$Z = \frac{(\Pi_1 - \Pi_2) - \delta_0}{\sqrt{\left(\frac{p_1 q_1}{n_1}\right) + \left(\frac{p_2 q_2}{n_2}\right)}} \sim N(0, 1) \quad (15.135)$$

Since population values,  $p_1$  and  $p_2$ , are seldom available in practice, it is customary to substitute sample estimates,

$$\hat{p}_1 = \frac{x_1}{n_1}; \text{ and } \hat{p}_2 = \frac{x_2}{n_2} \quad (15.136)$$

Finally, since this test statistic possesses a standard normal distribution, the rejection regions are precisely the same as those in Table 15.4.

In the special case when  $\delta_0 = 0$ , which is equivalent to a test of equality of the proportions, the most important consequence is that if the null hypothesis is true, then  $p_1 = p_2 = p$ , which is then estimated by the “pooled” proportion:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} \quad (15.137)$$

As a result, the standard deviation of the difference in proportions,  $\sigma_{D_{\Pi}}$ , becomes:

$$\sigma_{D_{\Pi}} = \sqrt{\left(\frac{p_1 q_1}{n_1}\right) + \left(\frac{p_2 q_2}{n_2}\right)} \approx \sqrt{\hat{p} \hat{q} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \quad (15.138)$$

so that the test statistic in Eq (15.135) is modified to

$$Z = \frac{(\Pi_1 - \Pi_2)}{\sqrt{\hat{p} \hat{q} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1) \quad (15.139)$$

The rejection regions are the same as in the general case.

**Example 15.14: REGIONAL PREFERENCE FOR PEPSI**

To confirm persistent rumors that the preference for PEPSI on engineering college campuses is higher in the Northeast of the United States than on comparable campuses in the Southeast, a survey was carried out on 125 engineering students chosen at random on the MIT campus in Cambridge, MA, and the same number of engineering students selected at random at Georgia Tech in Atlanta, GA. Each student was asked to indicate a preference for PEPSI versus other soft drinks, with the following results: 44 of the 125 at MIT indicate preference for PEPSI versus 26 at GA Tech. At the 5% level, determine whether the Northeast proportion,  $\hat{p}_1 = 0.352$ , is essentially the same as the Southeast proportion,  $\hat{p}_2 = 0.208$ , against the alternative that they are different.

**Solution:**

The hypotheses to be tested are:

$$\begin{aligned} H_0 : \Pi_1 - \Pi_2 &= 0 \\ H_a : \Pi_1 - \Pi_2 &\neq 0 \end{aligned} \quad (15.140)$$

and from the given data, the test statistic computed from Eq (15.139) is  $z = 2.54$ . Since this number is greater than 1.96, and therefore lies in the rejection region of the two-sided test, we reject the null hypothesis in favor of the alternative. Using MINITAB to carry out this test, selecting the "use pooled estimate of p for test," produces the following result:

**Test and CI for Two Proportions**

| Sample | X  | N   | Sample p |
|--------|----|-----|----------|
| 1      | 44 | 125 | 0.352000 |
| 2      | 26 | 125 | 0.208000 |

```
Difference = p (1) - p (2)
Estimate for difference: 0.144
95% CI for difference: (0.0341256, 0.253874)
Test for difference = 0 (vs not = 0): Z = 2.54 P-Value = 0.011
```

Note that the 95% confidence interval around the estimated difference of 0.144 does not include zero; the *p*-value associated with the test is 0.011 which is less than 0.05; hence, we reject the null hypothesis at the 5% significance level.

As an exercise, the reader should extend this example by testing  $\delta_0 = 0.02$  against the alternative that the difference is greater than 0.02.

## 15.8 Concerning Non-Gaussian Populations

The discussion in the previous section has opened up the issue of testing hypotheses about non-Gaussian populations, and has provided a strategy for handling such problems in general. The central issue is finding an appropriate test statistic and its sampling distribution, as was done for the binomial distribution. This cause is advanced greatly by the relationship between interval estimates and hypothesis tests (discussed earlier in Section 15.3.3) and by the discussion at the end of Chapter 14 on interval estimates for non-Gaussian distributions.

### 15.8.1 Large Sample Test for Means

First, if the statistical hypothesis is about the mean of a non-Gaussian population, so long as the sample size,  $n$ , used to compute the sample average,  $\bar{X}$ , is reasonably large (e.g.  $n > 30$  or so), then, regardless of the underlying distribution, we know that the statistic  $Z = (\bar{X} - \mu)/\sigma_{\bar{X}}$  possesses an approximate standard normal distribution — an approximation that improves as  $n \rightarrow \infty$ . Thus, hypotheses about the means of non-Gaussian populations that are based on large sample sizes are essentially the same as  $z$ -tests.

#### Example 15.15: HYPOTHESIS TEST ON MEAN OF INCLUSIONS DATA

If the data in Table 1.2 is considered a random sample of 60 observations of the number of *inclusions* found on glass sheets produced in the manufacturing process discussed in Chapter 1, test at the 5% significance level, the hypothesis that this data came from a Poisson population with mean  $\lambda = 1$ , against the alternative that  $\lambda$  is not 1.

#### Solution:

The hypotheses to be tested are:

$$\begin{aligned} H_0 : \lambda &= 1 \\ H_a : \lambda &\neq 1 \end{aligned} \quad (15.141)$$

While the data is from a Poisson population, the sample size is large; hence, the test statistic:

$$Z = \frac{\bar{X} - \lambda_0}{\sigma/\sqrt{60}} \quad (15.142)$$

where  $\sigma$  is the standard deviation of the raw data (so that  $\sigma/\sqrt{60}$  is the standard deviation of the sample average), essentially has a standard normal distribution.

From the supplied data, we obtain the sample average  $\hat{\lambda} = \bar{x} = 1.02$ , with the sample standard deviation,  $s = 1.1$ , which, because of the large sample, will be considered to be a reasonable approximation of  $\sigma$ . The

test statistic is therefore obtained as  $z = 0.141$ . Since this value is not in the two-sided rejection region  $|z| > 1.96$  for  $\alpha = 0.05$ , we do not reject the null hypothesis. We therefore conclude that there is no evidence to contradict the statement that  $X \sim \mathcal{P}(1)$ , i.e., the inclusions data is from a Poisson population with mean number of inclusions = 1.

It is now important to recall the results in Example 14.13 where the 95% confidence interval estimate for the mean of the inclusions data was obtained as:

$$\lambda = 1.02 \pm 1.96(1.1/\sqrt{60}) = 1.02 \pm 0.28 \quad (15.143)$$

i.e.,  $0.74 < \lambda < 1.30$ . Note that this interval contains the hypothesized value  $\lambda = 1.0$ , indicating that we cannot reject the null hypothesis.

We can now use this result to answer the following question raised in Chapter 1 as a result of the potentially “disturbing” data obtained from the quality control lab apparently indicating too many glass sheets with too many inclusions: *if the process was designed to produce glass sheets with a mean number of inclusions  $\lambda^* = 1$  per  $m^2$ , is there evidence in this sample data that the process has changed, that the number of observed “inclusions” is significantly different from what one can reasonably expect from the process when operating as designed?*

From the results of this example, the answer is, No: at the 5% significance level, there no evidence that the process has deviated from its design target.

### 15.8.2 Small Sample Tests

When the sample size on which the sample average is based is small, or when we are dealing with aspects of the population other than the mean, (say the variance), we are left with only one option: go back to “first principles,” derive the sampling distribution for the appropriate statistic and use it to carry out the required test. One can use the sampling distribution to determine  $\alpha \times 100\%$  rejection regions, or the complementary region, the  $(1 - \alpha) \times 100\%$  confidence interval estimates for the appropriate parameter.

For tests involving single parameters, it makes no difference which of these two approaches we choose; for tests involving two parameters, however, it is more straightforward to compute confidence intervals for the parameters in question and then use these for the hypothesis test. The reason is that for tests involving two parameters, confidence intervals can be computed directly from the individual sampling distributions; on the other hand, computing rejection regions for the difference between these two parameters technically requires an additional step of deriving yet another sampling distribution for the *difference*. And the sampling distribution of the difference between two random variables may not always be easy to derive. Having discussed earlier in this chapter the equivalence between confidence intervals and hypothesis tests, we now note that for non-Gaussian problems, one might as well just base the hypotheses tests on  $(1 - \alpha) \times 100\%$  confidence intervals and avoid the additional hassle of

having to derive distributions for differences. Let us illustrate this concept with a problem involving the exponential random variable discussed in Chapter 14.

In Example 14.3, we presented a problem involving an exponential random variable, the waiting time (in days) until the occurrence of a recordable safety incident in a certain company's manufacturing site. The safety performance data for the first and second years were presented, from which point estimates of the unknown population parameter,  $\beta$ , were determined from the sample averages,  $\bar{x}_1 = 30.1$  days, for Year 1 and  $\bar{x}_2 = 32.9$  days for Year 2; the sample size in each case is  $n = 10$ , which is considered small.

To test the two-sided hypothesis that these two safety performance parameters (Year 1 versus Year 2) are the same, versus the alternative that they are significantly different (at the 5% significance level), we proceed as follows: we first obtain the sampling distribution for  $\bar{X}_1$  and  $\bar{X}_2$  given that  $X \sim E(\beta)$ ; we then use these to obtain 95% confidence interval estimates for the population means  $\beta_i$  for Year  $i$ ; if these intervals overlap, then at the 5% significance level, we cannot reject the null hypothesis that these means are the same; if the intervals do not overlap, we reject the null hypothesis.

Much of this, of course, was already accomplished in Example 14.14: we showed that  $\bar{X}_i$  has a gamma distribution, more specifically,  $\bar{X}_i/\beta_i \sim \gamma(n, 1/n)$ , from where we obtain 95% confidence intervals estimates for  $\beta_i$  from sample data. In particular, for  $n = 10$ , we obtained from the  $\text{Gamma}(10, 0.1)$  distribution that:

$$P\left(0.48 < \frac{\bar{X}}{\beta} < 1.71\right) = 0.95 \quad (15.144)$$

which, upon introducing  $\bar{x}_1 = 30.1$ , and  $\bar{x}_2 = 32.9$ , produces, upon careful rearrangement, the 95% confidence interval estimates for the Year 1 and Year 2 parameters respectively as:

$$17.6 < \beta_1 < 62.71 \quad (15.145)$$

$$19.24 < \beta_2 < 68.54 \quad (15.146)$$

These intervals may now be used to answer a wide array of questions regarding hypotheses concerning two parameters, even questions concerning a single parameter. For instance,

1. For the two-parameter null hypothesis,  $H_0 : \beta_1 = \beta_2$ , versus  $H_a : \beta_1 \neq \beta_2$ , because the 95% confidence intervals overlap considerably, we find no evidence to reject  $H_0$  at the 5% significance level.
2. In addition, the single parameter null hypothesis,  $H_0 : \beta_1 = 40$ , versus  $H_a : \beta_1 \neq 40$ , cannot be rejected at the 5% significance level because the postulated value is contained in the 95% confidence interval for  $\beta_1$ ; on the contrary, the null hypothesis  $H_0 : \beta_1 = 15$ , versus  $H_a : \beta_1 \neq 15$  will be rejected at the 5% significance level because the hypothesized value falls outside of the 95% confidence interval (i.e., falls in the rejection region).

3. Similarly, the null hypothesis  $H_0 : \beta_2 = 40$ , versus  $H_a : \beta_2 \neq 40$ , cannot be rejected at the 5% significance level because the postulated value is contained in the 95% confidence interval for  $\beta_2$ ; on the other hand, the null hypothesis  $H_0 : \beta_2 = 17$ , versus  $H_a : \beta_2 \neq 17$  will be rejected at the 5% significance level — the hypothesized value falls outside of the 95% confidence interval (i.e., falls in the rejection region).

The principles illustrated here can be applied to any non-Gaussian population provided the sampling distribution of the statistic in question can be determined.

Another technique for dealing with populations characterized by any general pdf (Gaussian or not), and based on the maximum likelihood principle discussed in Chapter 14 for estimating unknown population parameters, is discussed next in its own separate section.

## 15.9 Likelihood Ratio Tests

In its broadest sense, a likelihood ratio (LR) test is a technique for assessing how well a simpler, “restricted” version of a probability model compares to its more complex, unrestricted version in explaining observed data. Within the context of this current chapter, however, the discussion here will be limited to testing hypotheses about the parameters,  $\boldsymbol{\theta}$ , of a population characterized by the pdf  $f(x, \boldsymbol{\theta})$ . Even though based on fundamentally different premises, some of the most popular tests considered above (the  $z$ - and  $t$ -tests, for example) are equivalent to LR tests under recognizable conditions.

### 15.9.1 General Principles

Let  $X$  be a random variable with the pdf  $f(x, \boldsymbol{\theta})$ , where the population parameter vector  $\boldsymbol{\theta} \in \Theta$ ; i.e.,  $\Theta$  represents the set of possible values that the parameter vector can take. Given a random sample,  $X_1, X_2, \dots, X_n$ , estimation theory, as discussed in Chapter 14, is concerned with using such sample information to determine reasonable estimates for  $\boldsymbol{\theta}$ . In particular, we recall that the maximum likelihood (ML) principle requires choosing the estimate,  $\hat{\boldsymbol{\theta}}_{ML}$ , as the value of  $\boldsymbol{\theta}$  that maximizes the likelihood function:

$$L(\boldsymbol{\theta}) = f_1(x_1, \boldsymbol{\theta})f_2(x_2, \boldsymbol{\theta}) \cdots f_n(x_n, \boldsymbol{\theta}) \quad (15.147)$$

the joint pdf of the random sample, treated as a function of the unknown population parameter.

The same random sample and the same ML principle can be used to test the null hypotheses

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \quad (15.148)$$

stated in a more general fashion in which  $\boldsymbol{\theta}$  is restricted to a certain range of values,  $\Theta_0$  (a subset of  $\Theta$ ), over which  $H_0$  is hypothesized to be valid. For example, to test a hypothesis about the mean of  $X$  by postulating that  $X \sim N(75, 1.5^2)$ , in this current context,  $\Theta$ , the full set of possible parameter values, is defined as follows:

$$\Theta = \{(\theta_1, \theta_2) : -\infty < \theta_1 = \mu < \infty; \theta_2 = \sigma^2 = 1.5^2\} \quad (15.149)$$

since the variance is given and the only unknown parameter is the mean;  $\Theta_0$ , the restricted parameter set range over which  $H_0$  is conjectured to be valid, is defined as:

$$\Theta_0 = \{(\theta_1, \theta_2) : \theta_1 = \mu_0 = 75; \theta_2 = \sigma^2 = 1.5^2\} \quad (15.150)$$

The null hypothesis in Eq (15.148) is to be tested against the alternative:

$$H_a : \boldsymbol{\theta} \in \Theta_a \quad (15.151)$$

again stated in a general fashion in which the parameter set,  $\Theta_a$ , is (a) disjoint from  $\Theta_0$ , and (b) also complementary to it, in the sense that

$$\Theta = \Theta_0 \cup \Theta_a \quad (15.152)$$

For example, the two-sided alternative to the hypothesis above regarding  $X \sim N(75, 1.5^2)$  translates to:

$$\Theta_a = \{(\theta_1, \theta_2) : \theta_1 = \mu_0 \neq 75; \theta_2 = \sigma^2 = 1.5^2\} \quad (15.153)$$

Note that the union of this set with  $\Theta_0$  in Eq (15.150) is the full parameter set range,  $\Theta$  in Eq (15.149).

Now, define the largest likelihood under  $H_0$  as

$$L^*(\Theta_0) = \max_{\boldsymbol{\theta} \in \Theta_0} L(\boldsymbol{\theta}) \quad (15.154)$$

and the unrestricted maximum likelihood value as:

$$L^*(\Theta) = \max_{\boldsymbol{\theta} \in \Theta_0 \cup \Theta_a} L(\boldsymbol{\theta}) \quad (15.155)$$

Then the ratio:

$$\Lambda = \frac{L^*(\Theta_0)}{L^*(\Theta)} \quad (15.156)$$

is known as the *likelihood ratio*; it possesses some characteristics that make it attractive for carrying out general hypothesis tests. But first, we note that by definition,  $L^*(\Theta)$  is the maximum value achieved by the likelihood function when  $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{ML}$ . Also,  $\Lambda$  is a random variable (it depends on the random sample,  $X_1, X_2, \dots, X_n$ ); this is why it is sometimes called the *likelihood ratio test statistic*. When specific data values,  $x_1, x_2, \dots, x_n$ , are introduced into Eq (15.156), the result is a specific value,  $\lambda$ , for the likelihood ratio such that  $0 \leq \lambda \leq 1$ , for the following reasons:

1.  $\lambda \geq 0$ . This is because each likelihood function contributing to the ratio is a pdf (joint pdfs, but pdfs nonetheless), and each legitimate pdf is such that  $f(x, \boldsymbol{\theta}) > 0$ ;
2.  $\lambda \leq 1$ . This is because  $\Theta_0 \subset \Theta$ ; consequently, since  $L^*(\Theta)$  is the largest achievable value of the likelihood function in the entire unrestricted set  $\Theta$ , the largest likelihood value achieved in the subset  $\Theta_0$ ,  $L^*(\Theta_0)$ , will be less than, or at best equal to,  $L^*(\Theta)$ .

Thus,  $\Lambda$  is a random variable defined on the unit interval  $(0,1)$  whose pdf,  $f(\lambda|\boldsymbol{\theta}_0)$  (determined by  $f(x, \boldsymbol{\theta})$ ), can be used, in principle, to test  $H_0$  in Eq (15.148) versus  $H_a$  in Eq (15.151). It should not come as a surprise that, in general, the form of  $f(\lambda|\boldsymbol{\theta}_0)$  can be quite complicated. However there are certain general principles regarding the use of  $\Lambda$  for hypothesis testing:

1. If a specific sample  $x_1, x_2, \dots, x_n$ , generates a value of  $\lambda$  close to zero, the implication is that the observation is highly unlikely to have occurred had  $H_0$  been true relative to the alternative;
2. Conversely, if  $\lambda$  is close to 1, then the likelihood of the observed data,  $x_1, x_2, \dots, x_n$ , occurring if  $H_0$  is true is just about as high as the unrestricted likelihood that  $\boldsymbol{\theta}$  can take any value in the entire unrestricted parameter space  $\Theta$ ;
3. Thus, small values of  $\lambda$  provide evidence *against* the validity of  $H_0$ ; larger values provide evidence in support.

How “small”  $\lambda$  has to be to trigger rejection of  $H_0$  is formally determined in the usual fashion: using the distribution for  $\Lambda$ , the pdf  $f(\lambda|\boldsymbol{\theta}_0)$ , obtain a critical value,  $\lambda_c$ , such that  $P(\Lambda < \lambda_c) = \alpha$ , i.e.,

$$P(\Lambda < \lambda_c) = \int_0^{\lambda_c} f(\lambda|\boldsymbol{\theta}_0) d\lambda = \alpha \quad (15.157)$$

Any value of  $\lambda$  less than this critical value will trigger rejection of  $H_0$ .

Likelihood ratio tests are very general; they can be used even for cases involving structurally different  $H_0$  and  $H_a$  probability distributions, or for random variables that are correlated. While the form of the pdf for  $\Lambda$  that is appropriate for each case may be quite complicated, in general, it is always possible to perform the required computations numerically using computer programs. Nevertheless, there are many special cases for which closed-form analytical expressions can be derived directly either for  $f(\lambda|\boldsymbol{\theta}_0)$ , the pdf of  $\Lambda$  itself, or else for the pdf of a monotonic function of  $\Lambda$ . See Pottmann *et al.*, (2005)<sup>2</sup>, for an application of the likelihood ratio test to an industrial sensor data analysis problem.

---

<sup>2</sup>Pottmann, M., B. A. Ogunnaike, and J. S. Schwaber, (2005). “Development and Implementation of a High-Performance Sensor System for an Industrial Polymer Reactor,” *Ind. Eng. Chem. Res.*, 44, 2606-2620.

### 15.9.2 Special Cases

#### Normal Population; Known Variance

Consider first the case where a random variable  $X \sim N(\mu, \sigma^2)$ , has known variance, but an unknown mean; and let  $X_1, X_2, \dots, X_n$  be a random sample from this population. From a specific sample data set,  $x_1, x_2, \dots, x_n$ , we wish to test  $H_0 : \mu = \mu_0$  against the alternative,  $H_a : \mu \neq \mu_0$ .

Observe that in this case, with  $\boldsymbol{\theta} = (\theta_1, \theta_2) = (\mu, \sigma^2)$ , the parameter spaces of interest are:

$$\Theta_0 = \{(\theta_1, \theta_2) : \theta_1 = \mu_0; \theta_2 = \sigma^2\} \quad (15.158)$$

and,

$$\Theta = \Theta_0 \cup \Theta_a = \{(\theta_1, \theta_2) : -\infty < \theta_1 = \mu < \infty; \theta_2 = \sigma^2\} \quad (15.159)$$

Since  $f(x, \boldsymbol{\theta})$  is Gaussian, the likelihood function, given the data, is

$$\begin{aligned} L(\mu, \sigma) &= \prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} \exp \left\{ \frac{-(x_i - \mu)^2}{2\sigma^2} \right\} \\ &= \left( \frac{1}{2\pi} \right)^{n/2} \frac{1}{\sigma^n} \exp \left\{ \frac{-\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2} \right\} \end{aligned} \quad (15.160)$$

This function is maximized (when  $\sigma^2$  is known) by the maximum likelihood estimator for  $\mu$ , the sample average,  $\bar{X}$ ; thus, the unrestricted maximum value,  $L^*(\Theta)$ , is obtained by introducing  $\bar{X}$  for  $\mu$  in Eq (15.160); i.e.,

$$L^*(\Theta) = \left( \frac{1}{2\pi} \right)^{n/2} \frac{1}{\sigma^n} \exp \left\{ \frac{-\sum_{i=1}^n (x_i - \bar{X})^2}{2\sigma^2} \right\} \quad (15.161)$$

On the other hand, the likelihood function, restricted to  $\boldsymbol{\theta} \in \Theta_0$  (i.e.,  $\mu = \mu_0$ ) is obtained by introducing  $\mu_0$  for  $\mu$  in Eq (15.160). Because, in terms of  $\mu$ , this function is now a constant, its maximum (in terms of  $\mu$ ) is given by:

$$L^*(\Theta_0) = \left( \frac{1}{2\pi} \right)^{n/2} \frac{1}{\sigma^n} \exp \left\{ \frac{-\sum_{i=1}^n (x_i - \mu_0)^2}{2\sigma^2} \right\} \quad (15.162)$$

From here, the likelihood ratio statistic is obtained as:

$$\Lambda = \frac{\exp \left\{ \frac{-\sum_{i=1}^n (x_i - \mu_0)^2}{2\sigma^2} \right\}}{\exp \left\{ \frac{-\sum_{i=1}^n (x_i - \bar{X})^2}{2\sigma^2} \right\}} \quad (15.163)$$

Upon rewriting  $(x_i - \mu_0)^2$  as  $[(x_i - \bar{X}) - (\bar{X} - \mu_0)]^2$  so that:

$$\sum_{i=1}^n (x_i - \mu_0)^2 = \sum_{i=1}^n (x_i - \bar{X})^2 + n(\bar{X} - \mu_0)^2 \quad (15.164)$$

and upon further simplification, the result is:

$$\Lambda = \exp \left\{ \frac{-n(\bar{X} - \mu_0)^2}{2\sigma^2} \right\} \quad (15.165)$$

To proceed from here, we need the pdf for the random variable,  $\Lambda$ ; but rather than confront this challenge directly, we observe that:

$$-2 \ln \Lambda = \left( \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right)^2 = Z^2 \quad (15.166)$$

where  $Z$ , of course, is the familiar  $z$ -test statistic

$$Z = \left( \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right) \quad (15.167)$$

with a standard normal distribution,  $N(0, 1)$ . Thus the random variable,  $\Psi = -2 \ln \Lambda$ , therefore has a  $\chi^2(1)$  distribution. From here it is now a straightforward exercise to obtain the rejection region in terms of not  $\Lambda$ , but  $\Psi = -2 \ln \Lambda$  (or  $Z^2$ ). For a significance level of  $\alpha = 0.05$ , we obtain from tail area probabilities of the  $\chi^2(1)$  distribution that

$$P(Z^2 \geq 3.84) = 0.05 \quad (15.168)$$

so that the null hypothesis is rejected when:

$$\frac{n(\bar{X} - \mu_0)^2}{\sigma^2} > 3.84 \quad (15.169)$$

Upon taking square roots, being careful to retain both positive as well as negative values, we obtain the familiar rejection conditions for the  $z$ -test:

$$\begin{aligned} \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} &< -1.96 \text{ or} \\ \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} &> 1.96 \end{aligned} \quad (15.170)$$

The LR test under these conditions is therefore exactly the same as the  $z$ -test.

### Normal Population; Unknown Variance

When the population variance is unknown for the test discussed above, some things change slightly. First, the parameter spaces become:

$$\Theta_0 = \{(\theta_1, \theta_2) : \theta_1 = \mu_0; \theta_2 = \sigma^2 > 0\} \quad (15.171)$$

along with,

$$\Theta = \Theta_0 \cup \Theta_a = \{(\theta_1, \theta_2) : -\infty < \theta_1 = \mu < \infty; \theta_2 = \sigma^2 > 0\} \quad (15.172)$$

The likelihood function remains the same:

$$L(\mu, \sigma) = \left( \frac{1}{2\pi} \right)^{n/2} \frac{1}{\sigma^n} \exp \left\{ \frac{-\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2} \right\}$$

but this time both parameters are unknown, even though the hypothesis test is on  $\mu$  alone. As a result, the function is maximized by the maximum likelihood estimators for both  $\mu$ , and  $\sigma^2$  — respectively, the sample average,  $\bar{X}$ , and,

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n}$$

as obtained in Chapter 14.

The unrestricted maximum value,  $L^*(\Theta)$ , in this case is obtained by introducing these ML estimators for the respective unknown parameters in Eq (15.160) and rearranging to obtain:

$$L^*(\Theta) = \left\{ \frac{n}{2\pi \sum_{i=1}^n (x_i - \bar{X})^2} \right\}^{n/2} e^{-n/2} \quad (15.173)$$

When the parameters are restricted to  $\theta \in \Theta_0$ , this time, the likelihood function is maximized, after substituting  $\mu = \mu_0$ , by the MLE for  $\sigma^2$ , so that the largest likelihood value is obtained as:

$$L^*(\Theta_0) = \left\{ \frac{n}{2\pi \sum_{i=1}^n (x_i - \mu_0)^2} \right\}^{n/2} e^{-n/2} \quad (15.174)$$

Thus, the likelihood ratio statistic becomes:

$$\Lambda = \left\{ \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{\sum_{i=1}^n (x_i - \mu_0)^2} \right\}^{n/2} \quad (15.175)$$

And upon employing the sum-of-squares identity in Eq (15.164), and simplifying, we obtain:

$$\Lambda = \left\{ \frac{1}{1 + \frac{n(\bar{X} - \mu_0)^2}{\sum_{i=1}^n (x_i - \bar{X})^2}} \right\}^{n/2} \quad (15.176)$$

If we now introduce the sample variance  $S^2 = \sum_{i=1}^n (x_i - \bar{X})^2 / (n - 1)$ , this expression is easily rearranged to obtain:

$$\Lambda = \left\{ \frac{1}{1 + \frac{1}{n-1} \frac{n(\bar{X} - \mu_0)^2}{S^2}} \right\}^{n/2} \quad (15.177)$$

As before, to proceed from here, we need to obtain the pdf for the random variable,  $\Lambda$ . However, once again, we recognize a familiar statistic embedded in Eq (15.177), i.e.,

$$T^2 = \left( \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \right)^2 \quad (15.178)$$

where  $T$  has the student's  $t$ -distribution with  $\nu = n - 1$  degrees of freedom. The implication therefore is that:

$$\Lambda^{2/\nu} = \frac{1}{1 + T^2/\nu} \quad (15.179)$$

From here we observe that because  $\Lambda^{2/\nu}$  (and hence  $\Lambda$ ) is a strictly monotonically decreasing function of  $T^2$  in Eq (15.179), then the rejection region  $\lambda < \lambda_c$  for which say  $P(\Lambda < \lambda_c) = \alpha$ , is exactly equivalent to a rejection region  $T^2 > t_c^2$ , for which,

$$P(T^2 > t_c^2) = \alpha \quad (15.180)$$

Once more, upon taking square roots, retaining both positive as well as negative values, we obtain the familiar rejection conditions for the  $t$ -test:

$$\begin{aligned} \frac{(\bar{X} - \mu_0)}{S/\sqrt{n}} &< -t_{\alpha/2}(\nu) \text{ or} \\ \frac{(\bar{X} - \mu_0)}{S/\sqrt{n}} &> t_{\alpha/2}(\nu) \end{aligned} \quad (15.181)$$

which, of course, is the one-sample, two-sided  $t$ -test for a normal population with unknown variance.

Similar results can be obtained for tests concerning the variance of a single normal population (yielding the  $\chi^2$ -test) or concerning two variances from independent normal populations, yielding the  $F$ -test.

The point, however, is that having shown that the LR tests in these well-known special cases reduce to tests with which we are already familiar, we have the confidence that in the more complicated cases, where the population pdfs are non-Gaussian and closed-form expressions for  $\Lambda$  cannot be obtained as easily, the results (mostly determined numerically), can be trusted.

### 15.9.3 Asymptotic Distribution for $\Lambda$

As noted repeatedly above, it is often impossible to obtain closed-form pdfs for the likelihood ratio test statistic,  $\Lambda$ , or for appropriate functions thereof. Nevertheless, for large samples, there exists an asymptotic distribution:

**Asymptotic Distribution Result for LR Test Statistic:** The distribution of the random variable  $\Psi = -2 \ln \Lambda$  tends asymptotically to a  $\chi^2(\nu)$  distribution with  $\nu$  degrees of freedom, with  $\nu = \mathcal{N}_p(\Theta) - \mathcal{N}_p(\Theta_0)$  where  $\mathcal{N}_p()$  is the number of independent parameters in the parameter space in question. i.e., the number of parameters in  $\Theta$  exceeds those in  $\Theta_0$  by  $\nu$ .

Observe, for example, that the distribution of  $\Psi = -2 \ln \Lambda$  in the first special case (Gaussian distribution with known variance) is exactly  $\chi^2(1)$ :  $\Theta$  contains one unknown parameter,  $\mu$ , while  $\Theta_0$  contains no unknown parameter since  $\mu = \mu_0$ .

This asymptotic result is exactly equivalent to the large sample approximation to the sampling distribution of means of arbitrary populations. Note that in the second special case (Gaussian distribution with *unknown* variance),  $\Theta$  contains two unknown parameter,  $\mu$  and  $\sigma^2$ , while  $\Theta_0$  contains only one unknown parameter,  $\sigma^2$ . The asymptotic distribution of  $\Psi = -2 \ln \Lambda$  will then also be  $\chi^2(1)$ , in precisely the same sense in which  $t(\nu) \rightarrow N(0, 1)$ .

---

## 15.10 Discussion

This chapter should not end without bringing to the reader's attention some of the criticisms of certain aspects of hypothesis testing. The primary issues have to do not so much with the mathematical foundations of the methodology as with the implementation and interpretation of the results in practice. Of several controversial issues, the following are three we wish to highlight:

1. *Point null hypothesis and statistical-versus-practical significance:* When the null hypothesis about a population parameter is that  $\theta = \theta_0$ , where  $\theta_0$  is a point on the real line, such a literal mathematical statement, can almost always be proven false with computations carried *to a sufficient number of decimal places*. For example, if  $\theta_0 = 75.5$ , a large enough sample that generates  $\bar{x} = 75.52$  (a routine possibility even when the population parameter is indeed 75.5) will lead to the rejection of  $H_0$ , to two decimal places. However, in actual practice (engineering or science), is the distinction between two real numbers 75.5 and 75.52 truly of importance? i.e., is the statement  $75.5 \neq 75.52$ , which is true in the strictest, literal mathematical sense, meaningful in practice? Sometimes yes, sometime no; but the point is that such null hypotheses can almost always be falsified, raising the question: what then does rejecting  $H_0$  really mean?
2. *Borderline p-values and variability:* Even when the *p*-value is used to determine whether or not to reject  $H_0$ , it is still customary to relate the computed *p*-value to some value of  $\alpha$ , typically 0.05. But what happens for  $p = 0.06$ , or  $p = 0.04$ ? Furthermore, an important fact that often goes unnoticed is that were we to repeat the experiment in question, the new data set will almost always lead to results that are "different" from those obtained earlier; and consequently the new *p*-value will also be different from that obtained earlier. One cannot therefore rule out the possibility of a "borderline" *p*-value "switching sides" purely as a result of intrinsic variability in the data.
3. *Probabilistic interpretations:* From a more technical perspective, if  $\delta$  represents the observed discrepancy between the observed postulated

population parameter and the value determined from data (a realization of the random variable,  $\Delta$ ), the  $p$ -value (or else the actual significance level of the test) is defined as  $P(\Delta \geq \delta | H_0)$ ; i.e., the probability of observing the computed difference or something more extreme if the null hypothesis is true. In fact, the probability we should be interested in is the reverse:  $P(H_0 | \Delta \geq \delta)$ , i.e., the probability that the null hypothesis is true given the evidence in the data, which truly measures how much the observed data supports the proposed statement of  $H_0$ . These two conditional probabilities are generally not the same.

In light of these issues (and others we have not discussed here), how should one approach hypothesis testing in practice? First, statistical significance should not be the only factor in drawing conclusions from experimental results — the nature of the problem at hand should be taken into consideration as well. The yield from process A may in fact not be precisely 75.5% (after all, the probability that a random variable will take on a precise value on the real line is exactly zero), but 75.52% is sufficiently close that the difference is of no practical consequence. Secondly, one should be careful in basing the entire decision about experimental results on a *single* hypothesis test, especially with  $p$ -values at the border of the traditional  $\alpha = 0.05$ . A single statistical hypothesis test of data obtained in a single study is just that: it can hardly be considered as having definitively “confirmed” something. Thirdly, decisions based on confidence intervals around the estimated population parameters tend to be less confusing and are more likely to provide the desired solution more directly.

Finally, the reader should be aware of the existence of other recently proposed alternatives to conventional hypothesis testing, e.g. Jones and Tukey (2000)<sup>3</sup>, or Killeen (2005)<sup>4</sup>. These techniques are designed to ameliorate some of the problems discussed above, but any discussions on them, even of the most cursory type, lie outside of the intended scope of this chapter. Although not yet as popular as the classical techniques discussed here, they are worth exploring by the curious reader.

### 15.11 Summary and Conclusions

If the heart of statistics is inference—drawing conclusions about populations from information in a sample—then this chapter and Chapter 14 jointly constitute the heart of Part IV of this book. Following the procedures dis-

<sup>3</sup>L. V. Jones, and J. W. Tukey, (2000): “A Sensible Formulation of the Significance Test,” *Psych. Methods*, 5 (4) 411-414

<sup>4</sup>P.R. Killeen, (2005): “An Alternative to Null-Hypothesis Significance Tests,” *Psychol Sci*, 16(5), 345-353

cussed in Chapter 14 for determining population parameters from sample data, we have focussed primarily in this chapter on procedures by which one makes and tests the validity of assertive statements about these population parameters. Thus, with some perspective, we may now observe the following: in order to characterize a population fully using the information contained in a finite sample drawn from it, (a) the results of Chapter 13 enable us to characterize the variability in the sample, so that (b) the unknown parameters may be estimated with a prescribed degree of confidence using the techniques in Chapter 14; and (c) what these estimated parameters tell us about the true population characteristics is then framed in the form of hypotheses that are subsequently tested using the techniques presented in this chapter. Specifically, the null hypothesis,  $H_0$ , is stated as the status quo characteristic; this is then tested against an appropriate alternative that we are willing to entertain should there be sufficient evidence in the sample data against the validity of the null hypothesis—each null hypothesis and the specific competing alternative having been jointly designed to answer the specific question of interest. This has been a long chapter, and perhaps justifiably so, considering the sheer number of topics covered; but the key results can be summarized briefly as we have done in Table 15.12 since hypotheses tests can be classified into a relatively small number of categories. There are tests for population *means* (for single populations or two populations; with population variance known, or not known; with large samples or small); there are also tests for (normal) population *variances* (single variances or two); and then there are tests for *proportions*, (one or two). In each case, once the appropriate test statistic is determined, with slight variations depending on specific circumstances, the principles are all the same. With fixed significance levels,  $\alpha$ , the  $H_0$  rejection regions are determined and are used straightforwardly to reach conclusions about each test. Alternatively, the *p*-value (also known as the *observed significance level*) is easily computed and used to reach conclusions. It bears restating that in carrying out the required computations not only in this chapter but in the book as a whole, we have consistently advocated the use of computer programs such as MINITAB. These programs are so widely available now that there is practically no need to make reference any longer to old-fashioned statistical tables. As a result, we have left out all but the most cursory references to any statistical tables, and instead included specific illustrations of how to use MINITAB (as an example software package).

The discussions of power and sample size considerations is important, both as a pre-experimentation design tool and as a post-analysis tool for ascertaining just how much stock one can realistically put in the result of a just-concluded test. Sadly, such considerations are usually given short-shrift by most students; this should *not* be the case. It is also easy to develop the mistaken notion that statistical inference is *only* concerned with Gaussian populations. Once more, as in Chapter 14, it is true that the *general* results we have presented have been limited to normal populations. This is due to the stubborn individuality of non-Gaussian distributions and the remarkable

versatility of the Gaussian distribution both in representing truly Gaussian populations (of course), but also as a reasonable approximation to the sampling distribution of the means of most non-Gaussian populations. Nevertheless, the discussion in Section 15.8 and the overview of likelihood ratio tests in Section 15.9 should serve to remind the reader that there is statistical inference life beyond samples from normal populations. A few of the exercises and application problems at the end of the chapter also buttress this point.

There is a sense in which the completion of this chapter can justifiably be considered as a pivotal point in the journey that began with the illustrative examples of Chapter 1. These problems, posed long ago in that introductory chapter, have now been fully solved in this chapter; and, in a very real sense, many practical problems can now be solved using only the techniques discussed up until this point. But this chapter is actually a convenient launching point for the rest of the discussion in this book, not a stopping point. For example, we have only discussed how to compare at most two population means; when the problem calls for the *simultaneous* comparison of more than two population means, the appropriate technique, ANOVA, is yet to be discussed. Although based on the *F*-test, to which we were introduced in this chapter, there is much more to the approach, as we shall see later, particularly in Chapter 19. Furthermore, ANOVA is only a part—albeit a foundational part—of Chapter 19, a chapter devoted to the design of experiments, the third pillar of statistics, which is concerned with ensuring that the samples used for statistical inference are as information rich as possible.

Immediately following this chapter, Chapter 16 (Regression Analysis) deals with estimation of a different kind, when the population parameters of interest are not constant as they have been thus far, but functions of another variable; naturally, much of the results of Chapter 14 and this current chapter are employed in dealing with such problems. Chapter 17 (Probability Model Validation) builds directly on the hypothesis testing results of this chapter in presenting techniques for explicitly validating postulated probability models; Chapter 18 (Nonparametric Methods) presents “distribution free” versions of many of the hypothesis tests discussed in this current chapter—a useful set of tools to have when one is unsure about the validity of the probability distributional assumptions (mostly the normality assumption) upon which classical tests are based. Even the remaining chapters beyond Chapter 19 (on case studies and special topics) all draw heavily from this chapter. A good grasp of the material in this chapter will therefore facilitate comprehension of the upcoming discussions in the remainder of the book.

---

## REVIEW QUESTIONS

- 1.** What is a statistical hypothesis?
- 2.** What differentiates a simple hypothesis from a composite one?
- 3.** What is  $H_0$ , the null hypothesis, and what is  $H_a$ , the alternative hypothesis?
- 4.** What is the difference between a two-sided and a one-sided hypothesis?
- 5.** What is a test of a statistical hypothesis?
- 6.** How is the US legal system illustrative of hypothesis testing?
- 7.** What is a test statistic?
- 8.** What is a critical/rejection region?
- 9.** What is the definition of the significance level of a hypothesis test?
- 10.** What are the types of errors to which hypothesis tests are susceptible, and what are their legal counterparts?
- 11.** What is the  $\alpha$ -risk, and what is the  $\beta$ -risk?
- 12.** What is the power of a hypothesis test, and how is it related to the  $\beta$ -risk?
- 13.** What is the sensitivity of a test as opposed to the specificity of a test?
- 14.** How are the performance measures, sensitivity and specificity, related to the  $\alpha$ -risk, and the  $\beta$ -risk?
- 15.** What is the  $p$ -value, and why is it referred to as the *observed significance level*?
- 16.** What is the general procedure for carrying out hypothesis testing?
- 17.** What test statistic is used for hypotheses concerning the single mean of a normal population when the variance is *known*?
- 18.** What is a  $z$ -test?
- 19.** What is an “upper-tailed” test as opposed to a “lower-tailed” test?
- 20.** What is the “one-sample”  $z$ -test?
- 21.** What test statistic is used for hypotheses concerning the single mean of a normal

population when the variance is *unknown*?

**22.** What is the “one-sample” *t*-test, and what differentiates it from the “one-sample” *z*-test?

**23.** How are confidence intervals related to hypothesis tests?

**24.** What test statistic is used for hypotheses concerning two normal population means when the variances are *known*?

**25.** What test statistic is used for hypotheses concerning two normal population means when the variances are *unknown* but equal?

**26.** What test statistic is used for hypotheses concerning two normal population means when the variances are *unknown* but unequal?

**27.** Is the distribution of the *t*-statistic used for the two-sample *t*-test with unknown and unequal variances an exact *t*-distribution?

**28.** What is a paired *t*-test, and what are the important characteristics that set the problem apart from the general two-sample *t*-test?

**29.** In determining power and sample size, what is the “*z*-shift”?

**30.** In determining power and sample size, what are the three hypothesis test characteristic parameters making up the “*z*-shift”? What is the equation relating them to the  $\alpha$ - and  $\beta$ -risks?

**31.** How can the  $\alpha$ -risk be reduced without simultaneously increasing the  $\beta$ -risk?

**32.** What are some practical considerations discussed in this chapter regarding the determination of the power of a hypothesis test and sample size?

**33.** For general power and sample size determination problems, it is typical to specify which two problem characteristics, leaving which three parameters to be determined?

**34.** What is the test concerning the single variance of a normal population variance called?

**35.** What test statistic is used for hypotheses concerning the single variance of a normal population?

**36.** What test statistic is used for hypotheses concerning two variances from mutually independent normal populations?

**37.** What is the *F*-test?

**38.** The *F*-test is quite sensitive to which assumption?

- 39.** What test statistic is used in the large sample approximation test concerning a single population proportion?
- 40.** What is the objective criterion for ascertaining the validity of the large sample assumption in tests concerning a single population proportion?
- 41.** What is involved in exact tests concerning a single population proportion?
- 42.** What test statistic is used for hypotheses concerning two population proportions?
- 43.** What is the central issue in testing hypotheses about non-Gaussian populations?
- 44.** How does sample size influence how hypotheses about non-Gaussian populations are tested?
- 45.** What options are available when testing hypotheses about non-Gaussian populations with small samples?
- 46.** What are likelihood ratio tests?
- 47.** What is the likelihood ratio test statistic?
- 48.** Why is the likelihood ratio parameter  $\lambda$  such that  $0 < \lambda < 1$ ? What does a value close to zero indicate? And what does a value close to 1 indicate?
- 49.** Under what condition does the likelihood ratio test become identical to the familiar  $z$ -test?
- 50.** Under what condition does the likelihood ratio test become identical to the familiar  $t$ -test?
- 51.** What is the asymptotic distribution result for the likelihood ratio statistic?
- 52.** What are some criticisms of hypothesis testing highlighted in this chapter?
- 53.** In light of some of the criticisms discussed in this chapter, what recommendations have been proposed for approaching hypothesis testing in practice?

---

## EXERCISES

### Section 15.2

**15.1** The target “mooney viscosity” of the elastomer produced in a commercial process is 44.0; if the average “mooney viscosity” of product samples acquired from the process hourly and analyzed in the quality control laboratory exceeds or falls

below this target, the process is deemed “out of control” and in need of corrective control action. Formulate the decision-making about the process performance as a hypothesis test, stating the null and the alternative hypotheses.

**15.2** A manufacturer of energy-saving light bulbs wants to establish that the lifetime of its new brand exceeds the specification of 1,000 hours. State the appropriate null and alternative hypotheses.

**15.3** A pharmaceutical company wishes to show that its newly developed acne medication reduces teenage acne by an average of 55% in the first week of usage. What are the null and alternative hypotheses?

**15.4** The owner of a fleet of taxi cabs wants to determine if there is a difference in the lifetime of two different brands of car batteries used in the fleet of cabs. State the appropriate null and alternative hypotheses.

**15.5** The safety coordinator of a manufacturing facility wishes to demonstrate that the mean time (in days) between safety incidents has deteriorated from the traditional 30 days. What are the appropriate null and alternative hypotheses?

**15.6** Consider  $X_1, X_2, \dots, X_n$ , a random sample from a normal population with a postulated mean  $\mu_0$  but known variance;

(i) If a test is based on the following criterion

$$P\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > 1.65\right)$$

what is (a) the type of hypothesis being tested; (ib) the test statistic; and (c) the significance level?

(ii) If instead, the variance is unknown, and the criterion is changed to:

$$P\left(\frac{\bar{X} - \mu_0}{s/\sqrt{n}} > 2.0\right) = 0.051$$

what is  $n$ ?

**15.7** Given a sample of size  $n = 15$  from a normal distribution with unknown mean and variance, a  $t$ -test statistic of 2.0 was determined for a one-sided, upper-tailed test. Determine the associated  $p$ -value. From a different sample of unknown size drawn from the same normal population, the rejection region for a test at the significance level of  $\alpha = 0.05$  was determined as  $t > 1.70$ . What is  $n$ ?

**15.8** Consider a random sample,  $X_1, X_2, \dots, X_n$  from an exponential population,  $\mathcal{E}(\beta)$ . It is known that the sample mean,  $\bar{X}$ , possesses a gamma  $\gamma(n, \beta/n)$  distribution. A hypothesis test regarding the population parameter  $\beta$  is to be determined using the following criterion:

$$P\left(C_L < \frac{\bar{X}}{\beta} < C_R\right) = 0.95$$

For this problem, what is (i) the type of hypothesis being tested; (ii) the test statistic; (iii) the significance level; and (iv) if  $n = 20$ , the rejection region?

**Section 15.3**

**15.9** A random sample of size  $n = 16$  drawn from a normal population with hypothesized mean  $\mu_0 = 50$  and known variance,  $\sigma^2 = 25$ , produced a sample average  $\bar{x} = 47.5$ .

- (i) Compute the appropriate test statistic.
- (ii) If the alternative hypothesis is  $H_a : \mu < \mu_0$ , at the  $\alpha = 0.1$  significance level, determine the rejection region. Should the null hypothesis  $H_0 : \mu = \mu_0$  be rejected?

**15.10** Refer to Exercise 15.9. Determine the 95% confidence interval estimate for the population mean and compare it with the hypothesized mean. What does this imply about whether or not the null hypothesis  $H_0 : \mu = \mu_0$  should be rejected at the  $\alpha = 0.05$  level? Determine the  $p$ -value associated with this test.

**15.11** Refer to Exercise 15.9. If instead the population variance is unknown and the sample variance is obtained as  $s^2 = 39.06$ , should the null hypothesis  $H_0 : \mu = \mu_0$  be rejected at the  $\alpha = 0.1$  level, and the  $\alpha = 0.05$  level?

**15.12.** The following random sample was obtained from a normal population with variance given as 1.00.

$$S_N = \{9.37, 8.86, 11.49, 9.57, 9.15, 9.10, 10.26, 9.87, 7.82, 10.47\}$$

To test the hypothesis that the population mean is different from a postulated value of 10.00, (i) state the appropriate null and alternative hypotheses; (ii) determine the appropriate test statistic; (iii) determine the rejection region for a hypothesis test at the  $\alpha = 0.05$  significance level; (iv) determine the associated  $p$ -value. (v) What conclusion should be drawn from this test?

**15.13** Refer to Exercise 15.12. Repeat for the case where the population variance is unknown. Does this fact change the conclusion of the test?

**15.14** A random sample of size  $n = 50$  from a normal population with  $\sigma = 3.00$  produced a sample mean of 80.05. At a significance level  $\alpha = 0.05$ ,

- (i) Test the null hypothesis that the population mean  $\mu_0 = 75.00$  against the alternative that  $\mu_0 > 75.00$ ; interpret the result of the test.
- (ii) Test the null hypothesis against the alternative  $\mu_0 \neq 75.00$ . Interpret the result of this test and compare it to the test in (i). Why are the results different?

**15.15** In carrying out a hypothesis test of  $H_0 : \mu = 100$  versus the alternative,  $H_a : \mu > 100$ , given that the population variance is 1600, it has been recommended to “reject  $H_0$  in favor of  $H_a$  if the mean of a random sample of size  $n = 100$  exceeds 106.” What is  $\alpha$ , the significance level behind the statement?

**Section 15.4**

**15.16** Random samples of 50 observations are each drawn from two independent normal populations,  $N(10.0, 2.5^2)$ , and  $N(12.0, 3.0^2)$ ; if  $\bar{X}$  represents the sample mean from the first population, and  $\bar{Y}$  is the sample mean from the second population,

- (i) Determine the sampling distribution for  $\bar{X}$  and  $\bar{Y}$ ;
- (ii) Determine the mean and variance of the sampling distribution for  $\bar{Y} - \bar{X}$ ;

- (iii) Determine the  $z$ -statistic associated with actual sample averages obtained as  $\bar{x} = 10.9$  and  $\bar{y} = 11.8$ . Use this to test  $H_0 : \mu_Y = \mu_X$  against  $H_a : \mu_Y > \mu_X$ .  
 (iv) Since we know from the supplied population information that  $\mu_Y > \mu_X$ , interpret the results of the test in (iii).

**15.17** Two samples of sizes  $n_1 = 20$  and  $n_2 = 15$  taken from two independent normal populations with known standard deviations,  $\sigma_1 = 3.5$  and  $\sigma_2 = 4.2$ , produced sample averages,  $\bar{x}_1 = 15.5$  and  $\bar{x}_2 = 13.8$ . At the  $\alpha = 0.05$  significance level, test the null hypothesis that the means are equal against the alternative that they are not. Interpret the result. Repeat this test for  $H_a : \mu_1 > \mu_2$ ; interpret the result.

**15.18** The data in the table below is a random sample of 15 observations each from two normal populations with unknown means and variances. Test the null hypothesis that the two population means are equal against the alternative that  $\mu_Y > \mu_X$ . First assume that the two population variances are equal. Interpret your results. Repeat the test without assuming equal variances. Is there a difference in the conclusions?

| Sample | X     | Y     |
|--------|-------|-------|
| 1      | 12.03 | 13.74 |
| 2      | 13.01 | 13.59 |
| 3      | 9.75  | 10.75 |
| 4      | 11.03 | 12.95 |
| 5      | 5.81  | 7.12  |
| 6      | 9.28  | 11.38 |
| 7      | 7.63  | 8.69  |
| 8      | 5.70  | 6.39  |
| 9      | 11.75 | 12.01 |
| 10     | 6.28  | 7.15  |
| 11     | 12.53 | 13.47 |
| 12     | 10.22 | 11.57 |
| 13     | 7.17  | 8.81  |
| 14     | 11.36 | 13.10 |
| 15     | 9.16  | 11.32 |

**15.19** Refer to Exercise 15.18. (i) On the same graph, plot the data for  $X$  and for  $Y$  against sample number. Comment on any feature that might indicate whether the two samples can be treated as independent or not.

(ii) Treat the samples as 15 paired observations and test the null hypothesis that the two population means are equal against the alternative that  $\mu_Y > \mu_X$ . Interpret your result and compare it with the results of Exercise 15.18.

**15.20** The data below are random samples from two independent lognormal distributions; specifically,  $X_{L_1} \sim \mathcal{L}(0, 0.25)$  and  $X_{L_2} \sim \mathcal{L}(0.25, 0.25)$ .

| $X_{L_1}$ | $X_{L_2}$ |
|-----------|-----------|
| 0.81693   | 1.61889   |
| 0.96201   | 1.15897   |
| 1.03327   | 1.17163   |
| 0.84046   | 1.09065   |
| 1.06731   | 1.27686   |
| 1.34118   | 0.91838   |
| 0.77619   | 1.45123   |
| 1.14027   | 1.47800   |
| 1.27021   | 2.16068   |
| 1.69466   | 1.46116   |

- (i) For the time being, ignore the fact that the sample size is too small to make the normal approximation valid for the sampling distribution of the sample means. At the  $\alpha = 0.05$  significance level, carry out a two-sample  $t$ -test concerning the equality of the means of these two populations, against the alternative that they are not equal. Interpret your results in light of the fact that we know that the two populations are *not* equal.
- (ii) Fortunately, a logarithmic transformation of lognormally distributed data yields normally distributed data; as a result, let  $Y_1 = \ln X_{L_1}$  and  $Y_2 = \ln X_{L_2}$  and repeat (i) for the log transformed  $Y_1$  and  $Y_2$  data. Interpret your results.
- (iii) Comment on the implication of these results on the inappropriate use of the normal approximation as well as the use of  $\alpha = 0.05$  in a dogmatic fashion.

**15.21** The sample averages  $\bar{X} = 38.8$  and  $\bar{Y} = 42.4$  were obtained from random samples taken from two independent populations of respective sizes  $n_x = 120$  and  $n_y = 90$ . The corresponding sample standard deviations were obtained as  $s_x^2 = 20$ ;  $s_y^2 = 35$ . At the  $\alpha = 0.05$  significance level, test the hypothesis that the population mean  $\mu_Y$  is greater than  $\mu_X$ . Interpret your result. How will the result change if instead, the hypothesis to be tested is that the two population means are different?

### Section 15.5

**15.22** A random sample of size  $n = 100$  from a normal population with unknown mean and variance is to be used to test the null hypothesis,  $H_0 : \mu = 12$ , versus the alternative,  $H_0 : \mu \neq 12$ . The observed sample standard deviation is  $s = 0.5$ .

- (i) Determine the rejection region for  $\alpha = 0.1, 0.05$  and  $\alpha = 0.01$ .
- (ii) If the true population mean has shifted to  $\mu = 11.9$ , determine the value of  $\beta$  corresponding to each of the rejection regions obtained in (i) and hence the power of each test. Comment on the effect that lowering  $\alpha$  has on the corresponding values of  $\beta$  and power.

**15.23** In the following, given  $\alpha = 0.05$  and  $\sigma = 1.0$ , determine the missing hypothesis test characteristic parameter for a two-sided, 1-sample  $z$ -test:

- (i) Power = 0.9, sample size = 40; and Power = 0.9, sample size = 20.
- (ii) Power = 0.75, sample size = 40; and Power = 0.75, sample size = 20.
- (iii) Power = 0.9, hypothesized mean shift to be detected = 0.5, and Power = 0.9, hypothesized mean shift to be detected = 0.75.
- (iv) Power = 0.75, hypothesized mean shift to be detected = 0.5, and Power = 0.75, hypothesized mean shift to be detected = 0.75.
- (v) Hypothesized mean shift to be detected = 0.5, sample size = 40; and Hypothe-

sized mean shift to be detected = 0.5, sample size = 20.

(vi) Hypothesized mean shift to be detected = 0.75, sample size = 40; and Hypothesized mean shift to be detected = 0.75, sample size = 20.

**15.24** Refer to Exercise 15.23. Repeat for 1-sample *t*-test. Comment on any differences in the computed characteristic parameter and whether or not this difference will mean anything in practice.

**15.25** Refer to the data in Exercise 15.20 where a logarithmic transformation of the data yielded  $Y_1$  and  $Y_2$  that are random samples from normal populations. The respective postulated means are 0 and 0.25, and the postulated standard deviations are both equal to 0.25. What is the power of the two-sided test of  $H_0 : \mu_{Y_1} = \mu_{Y_2} = 0$  versus  $H_a : \mu_{Y_1} \neq \mu_{Y_2}$  (because  $\mu_{Y_2} = 0.25$ ), carried out with the indicated sample of 10 observations at the  $\alpha = 0.05$  significance level? What sample size would have been required in order to carry out this test with power 0.9 or better?

**15.26** Samples are to be taken from two independent normal populations, one with variance  $\sigma_1^2 = 10$ , the other with a variance twice the magnitude of the first one. If the difference between the two population means is to be estimated to within  $\pm 2$  with a two-sample test at the  $\alpha = 0.05$  significance level, determine the sample size required for the test to have power of 0.9 or better, assuming that  $n_1 = n_2 = n$ . State any other assumptions needed to answer this question.

**15.27** If two variables  $y$  and  $x$  are related according to

$$y = f(x)$$

the relative sensitivity of  $y$  to changes in  $x$  is defined as:

$$S_r = \frac{\partial \ln y}{\partial \ln x}$$

It provides a measure of how relative changes  $\Delta x/x$  in  $x$  translate to corresponding relative changes  $\Delta y/y$  in  $y$ .

Now recall the expression given in Eq (15.98) which relates the sample size required to detect at a minimum, a signal of magnitude  $\delta^*$ , in the midst of intrinsic noise, characterized by standard deviation  $\sigma$ , using an upper-tailed test at the significance level  $\alpha = 0.05$  and with power  $(1 - \beta) = 0.9$ , i.e.,

$$n = \left( \frac{2.925}{\rho_{SN}} \right)^2$$

Show that the relative sensitivity of the sample size  $n$  to the signal-to-noise ratio  $\rho_{SN} = \delta^*/\sigma$  is  $-2$ , thus establishing that a 1% increase in the signal-to-noise ratio  $\rho_{SN}$  translates to an (instantaneous) incremental reduction of 2% in sample size requirements. Comment on ways by which one might increase signal-to-noise ratio in practical problems.

### Section 15.6

**15.28** The variance of a sample of size  $n = 20$  drawn from a normal population with mean 100 was obtained as  $s^2 = 9.5$ . At the  $\alpha = 0.05$  significance level, test the

hypothesis that the true population variance is 10.

**15.29** Refer to the data in Exercise 15.20. A logarithmic transformation of the data is postulated to yield  $Y_1$  and  $Y_2$ , random samples from normal populations with respective postulated means 0 and 0.25, and postulated equal standard deviation 0.25. Is there evidence in the log-transformed data to support the hypothesis that  $\sigma_{Y_1} = 0.25$ , and the hypothesis that  $\sigma_{Y_2} = 0.25$ ?

**15.30** A sample of 20 observations from a normal population was used to carry out a test concerning the unknown population variance at the  $\alpha = 0.05$  significance level. The hypothesis that the population variance is equal to a postulated value,  $\sigma_0$ , was eventually rejected in favor of the alternative that the population variance is higher. What is the relationship between the observed sample variance  $s^2$  and the postulated variance?

**15.31** Refer to Exercise 15.12 and the supplied data purported to be a random sample obtained from a normal population with variance given as 1.00.

$$S_N = \{9.37, 8.86, 11.49, 9.57, 9.15, 9.10, 10.26, 9.87, 7.82, 10.47\}$$

At the  $\alpha = 0.05$  significance level confirm or refute this postulate about the population variance.

**15.32** Refer to Exercise 15.29 and confirm directly the postulate that the two variances,  $\sigma_{Y_1}$  and  $\sigma_{Y_2}$  are equal. State the  $p$ -value associated with the test.

**15.33** (i) Determine the rejection region to be used in testing  $H_0 : \sigma_1^2 = \sigma_2^2$  against  $H_a : \sigma_1^2 \neq \sigma_2^2$ , with  $\nu_1 = 30, \nu_2 = 25$ , for each of the following cases: (a)  $\alpha = 0.05$  (b)  $\alpha = 0.10$

(ii) Repeat (i) when the alternative is  $H_a : \sigma_1^2 > \sigma_2^2$

**15.34** A random sample of size  $n_1 = 12$  from a normal population with unknown mean  $\mu_1$ , and unknown variance  $\sigma_1^2$ , and another independent random sample of size  $n_2 = 13$  from a different normal population with unknown mean  $\mu_2$ , and unknown variance  $\sigma_2^2$ , are to be sued to test the null hypothesis  $H_0 : \sigma_2^2 = k\sigma_1^2$  against the alternative  $H_a : \sigma_2^2 > k\sigma_1^2$ . Using the  $F$ -statistic,  $S_2^2/S_1^2$ , the critical region was obtained as  $f > 5.58$  at the  $\alpha = 0.05$  significance level. Determine the value of the constant  $k$ .

**15.35** The risk associated with two different stocks is quantified by the “volatility” i.e., the variability in daily prices was determined as  $s_1^2 = 0.58$  for the first one and  $s_2^2 = 0.21$  from the second, the variances having been determined from a random sample of 25 daily price changes in each case. Test the hypothesis  $H_0 : \sigma_1^2 = \sigma_2^2$  against  $H_a : \sigma_1^2 > \sigma_2^2$ . What is the  $p$ -value of the test? Interpret your result.

### Section 15.7

**15.36** A random sample of 100 observations from a binomial population resulted in an estimate  $\hat{p} = 0.72$  of the true population proportion,  $p$ . Is the sample size large enough to use the large sample approximation to test the hypothesis  $H_0 : p = 0.75$  against the alternative  $H_a : p \neq 0.75$ ? Test the indicated hypothesis at the  $\alpha = 0.05$

level and interpret your result. What is the associated  $p$ -value?

**15.37** A random sample of 50 observations from a binomial population was used to estimate the population parameter  $p$  as  $\hat{p} = 0.63$ .

- (i) Construct a 95% confidence interval for  $p$ .
- (ii) If it is desired to obtain an estimate of  $p$  to within  $\pm 0.10$  with 95% confidence, what sample size would be required?
- (iii) Extend the result in (ii) to the general case: determine an expression for the sample size required to estimate the population parameter  $p$  with  $\hat{p}$ , with the desired limits of the  $(1 - \alpha) \times 100\%$  confidence interval specified as  $\pm \ell$ .

**15.38** A supposedly fair coin was tossed 10 times and 4 heads were obtained. Test the hypothesis that the coin is fair versus the alternative that it is not fair. Use both the large sample approximation (check that this is valid first) and the exact test. Compare the two  $p$ -values associated with each test and interpret your result. Had the coin been tossed only 5 times with 2 heads resulting (so that the same  $\hat{p}$  would have been obtained), what will change in how the hypothesis test is carried out?

**15.39** A random sample of  $n_1 = 100$  observations from one binomial population and a separate random sample of  $n_2 = 75$  observations from a second binomial population produced  $x_1 = 33$  and  $x_2 = 27$  successes respectively.

- (i) At the  $\alpha = 0.05$  significance level, test the hypothesis that the two population proportions are equal against the alternative that they are different.
- (ii) Repeat (i) for the alternative hypothesis that  $p_2 > p_1$ .

**15.40** Two binomial population proportions are suspected to be approximately 0.3, but they are unknown precisely. Using these as initial estimates, it is desired to conduct a study to determine the difference between these two population proportions. Obtain a general expression for the sample size  $n_1 = n_2 = n$  required in order to determine the difference  $p_1 - p_2$  to within  $\pm \ell$  with  $(1 - \alpha) \times 100\%$  confidence. In the specific case where  $\ell = 0.02$  and  $\alpha = 0.05$  what is  $n$ ?

### Section 15.8 and 15.9

**15.41** The following data is sampled from an exponential population with unknown parameter  $\beta$ .

|       |      |      |      |       |
|-------|------|------|------|-------|
| 6.99  | 2.84 | 0.41 | 3.75 | 2.16  |
| 0.52  | 0.67 | 2.72 | 5.22 | 16.65 |
| 10.36 | 1.66 | 3.26 | 1.78 | 1.31  |
| 5.75  | 0.12 | 6.51 | 4.05 | 1.52  |

- (i) On the basis of the exact sampling distribution of the sample mean,  $\bar{X}$ , determine a 95% confidence interval estimate of the population parameter,  $\beta$ .
- (ii) Test the hypothesis  $H_0 : \beta = 4$  versus the alternative  $H_0 : \beta \neq 4$ . What is the  $p$ -value associated with this test?
- (iii) Repeat the test in (ii) using the normal approximation (which is not necessarily valid in this case). Compare this test result with the one in (ii).

**15.42** Refer to Exercise 15.41. Using the normal approximation, test the hypothesis that the sample variance is 16 versus the alternative that it is not, at the  $\alpha = 0.05$

significance level. Comment on whether or not this result should be considered as reliable.

**15.43** Given a random sample,  $X_1, X_2, \dots, X_n$  from a gamma  $\gamma(\alpha, \beta)$  distribution, use the reproductive properties result discussed in Chapter 8 to obtain the sampling distribution of the sample mean,  $\bar{X} = (\sum_{i=1}^n X_i)/n$ . In the specific case where the population parameters are postulated to be  $\alpha = 2$  and  $\beta = 20$  so that the population mean is  $\mu = 40$ , a random sample of size  $n = 20$  yielded an average  $\bar{x} = 45.6$ . Obtain an exact 95% confidence interval estimate of the true population mean, and from this test, at the  $\alpha = 0.05$  significance level, the null hypothesis  $H_0 : \mu = 40$  versus the alternative,  $H_a : \mu \neq 40$ .

**15.44** Let  $X_1, X_2, \dots, X_n$  be a random sample from a normal  $N(\mu, \sigma^2)$  population with unknown mean and variance. With the parameters represented as  $\theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$ , use the likelihood ratio method to construct a hypothesis test for  $H_0 : \sigma^2 = \sigma_0^2$ . First obtain the likelihood ratio  $\Lambda$  and then show that instead of  $\Lambda$ , the test statistic

$$\Lambda' = \frac{(n-1)S^2}{\sigma_0^2}$$

should be used, where  $S^2$  is the sample variance. Hence, establish that the likelihood ratio test for the variance of a single normal population is identical to the result obtained in Section 15.6.

**15.45** Let  $X_1, X_2, \dots, X_n$  be a random sample from an exponential population  $\mathcal{E}(\beta)$ . Establish that the likelihood ratio test of the hypothesis that  $\beta = \beta_0$  versus the alternative that  $\beta \neq \beta_0$  will result in a rejection region obtained from the solution to the following inequality:

$$\frac{\bar{x}}{\beta_0} e^{-\bar{x}/\beta_0} \leq k$$

where  $\bar{x}$  is the observed sample average, and  $k$  is a constant.

## APPLICATION PROBLEMS

**15.46** In a study to determine the performance of processing machines used to add raisins to trial-size “Raisin Bran” cereal boxes, (see Example 12.3 in Chapter 12), 6 sample boxes are taken at random from each machine’s production line and the number of raisins in each box counted. The result for machines 3 and 4 are shown below. Assume that these can be considered as random samples from a normal population.

| Machine 3 | Machine 4 |
|-----------|-----------|
| 13        | 7         |
| 7         | 4         |
| 11        | 7         |
| 9         | 7         |
| 12        | 12        |
| 18        | 18        |

- (i) If the target average number of raisins dispensed per box is 10, by carrying out appropriate hypothesis tests determine which of the two machines is operating according to target and which is not. State the  $p$ -value associated with each test.  
(ii) Is there any significant difference between the mean number of raisins dispensed by these two machines? Support your answer adequately.

**15.47** The data table below shows the wall thickness (in ins) of cast aluminum cylinder heads used in aircraft engine cooling jackets, taken from Mee (1990)<sup>5</sup>.

|       |       |       |       |       |       |       |       |       |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 0.223 | 0.228 | 0.214 | 0.193 | 0.223 | 0.213 | 0.218 | 0.215 | 0.233 |
| 0.201 | 0.223 | 0.224 | 0.231 | 0.237 | 0.217 | 0.204 | 0.226 | 0.219 |

If the manufacturing process is designed to produce cylinder heads whose wall thicknesses follow a normal distribution with mean wall thickness of 0.22 ins and standard deviation 0.01 ins, confirm or refute the claim that the process was operating as designed when the samples shown in the table were obtained. State any assumptions you may need to make in answering this question and keep in mind that there are *two* separate parameters in the postulated process characteristics.

**15.48** The data set below,  $S_{10}$ , from Holmes and Mergen (1992)<sup>6</sup>, is a sample of viscosity measurements taken from ten consecutive, but independent, batches of a product made in a batch chemical process.

$$S_{10} = \{13.3, 14.5, 15.3, 15.3, 14.3, 14.8, 15.2, 14.9, 14.6, 14.1\}$$

The desired target value for the product viscosity is 14.9. Assuming that the viscosity data constitutes a random sample from a normal population with unknown mean and unknown variance, at the  $\alpha = 0.05$  significance level, test the hypothesis that the mean product viscosity is on target versus the alternative that it is not. What is the  $p$ -value associated with this test? Interpret your result. If the test were to be conducted at the  $\alpha = 0.10$  significance level, will this change your conclusion?

**15.49** Kerkhof and Geboers, (2005)<sup>7</sup>, presented a new approach to modeling multicomponent transport that is purported to yield more accurate predictions. To demonstrate the performance of their modeling approach, the authors determined, experimentally, the viscosity ( $10^{-5} Pa.s$ ) of 12 different gas mixtures and compared them to the corresponding values predicted by the classical Hirschfelder-Curtiss-Bird

<sup>5</sup>Mee, R. W., (1990). "An improved procedure for screening based on a correlated, normally distributed variable," *Technometrics*, 32, 331–337.

<sup>6</sup>Holmes, D.S., and A.E. Mergen, (1992). "Parabolic control limits for the exponentially weighted moving average control charts" *Qual. Eng.* 4, 487–495.

<sup>7</sup>Kerkhof, P.J.A.M, and M.A.M. Geboers, (2005). "Toward a unified theory of isotropic molecular transport phenomena," *AICHE Journal*, 51, (1), 79–121

(HCB) model<sup>8</sup> and their new (KG) model. The results are shown in the table below.

| Viscosity, ( $10^{-5} \text{ Pa.s}$ ) |                 |                |
|---------------------------------------|-----------------|----------------|
| Experimental Data                     | HCB Predictions | KG Predictions |
| 2.740                                 | 2.718           | 2.736          |
| 2.569                                 | 2.562           | 2.575          |
| 2.411                                 | 2.429           | 2.432          |
| 2.504                                 | 2.500           | 2.512          |
| 3.237                                 | 3.205           | 3.233          |
| 3.044                                 | 3.025           | 3.050          |
| 2.886                                 | 2.895           | 2.910          |
| 2.957                                 | 2.938           | 2.965          |
| 3.790                                 | 3.752           | 3.792          |
| 3.574                                 | 3.551           | 3.582          |
| 3.415                                 | 3.425           | 3.439          |
| 3.470                                 | 3.449           | 3.476          |

- (i) Treated as paired data, perform an appropriate hypothesis test to compare the new KG model predictions with corresponding experimental results. Is there evidence to support the claim that this model provides “excellent agreement” with experimental data?
- (ii) Treated as paired “data,” test whether there is any significant difference between the HCB model predictions and the new KG model predictions.
- (iii) As in (i) perform a test to assess the performance of the classic HCB model prediction against experimental data. Can the HCB model be considered as also providing “excellent agreement” with experimental data?

**15.50** The table below, from Lucas (1985)<sup>9</sup>, shows the number of accidents occurring per quarter (three months), over a 10-year period, at a DuPont company facility. The data set is divided into two periods: Period I for the first five-year period of the study; Period II, for the second five-year period.

| Period I |   |    |    | Period II |   |   |   |
|----------|---|----|----|-----------|---|---|---|
| 5        | 5 | 10 | 8  | 3         | 4 | 2 | 0 |
| 4        | 5 | 7  | 3  | 1         | 3 | 2 | 2 |
| 2        | 8 | 6  | 9  | 7         | 7 | 1 | 4 |
| 5        | 6 | 5  | 10 | 1         | 2 | 2 | 1 |
| 6        | 3 | 3  | 10 | 4         | 4 | 4 | 4 |

- (i) Perform appropriate tests to confirm or refute the hypothesis that the true population mean number of accidents in the first period is 6, while the same population parameter was halved in the second period.
- (ii) Separately test the hypothesis that there is no significant difference between the mean number of accidents in each period. State any assumptions needed to answer these questions.

<sup>8</sup>Hirschfelder J.O., C.F. Curtiss, and R.B. Bird (1964). *Molecular Theory of Gases and Liquids*. 2nd printing. J. Wiley, New York, NY.

<sup>9</sup>Lucas J. M., (1985). “Counted Data CUSUMs,” *Technometrics*, 27, 129–144.

**15.51** A survey of alumni that graduated between 2000 and 2005 from the chemical engineering department of a University in the Mid-Atlantic region of the US involved 150 randomly selected individuals: 100 BS graduates and 50 MS graduates. (The PhD graduates participated in a different survey.) The survey showed, among other things, that 9.5% of BS graduates and 4.5% of MS graduates were unemployed for at least one year during this period.

- (i) If the corresponding national unemployment averages for *all* BS and MS degree holders in all engineering disciplines over the same period are, respectively, 15.2% and 7.5%, perform appropriate hypothesis tests to determine whether or not the chemical engineering alumni of this University fare better in general than graduates with corresponding degrees in other engineering disciplines.
- (ii) Does having an advanced degree make any difference in the unemployment status of the alumni of this University? Support your answer adequately.
- (iii) In connection with (ii) above, if it is desired to determine any true difference between the unemployment status of this University's alumni to within  $\pm 0.5\%$  with 95% confidence, how many alumni would have to be sampled? State any assumptions clearly.

**15.52** The data set in Problems 1.13 and 14.42, shown in the table below for ease of reference, is the time (in months) from receipt to publication of 85 papers published in the January 2004 issue of a leading chemical engineering research journal.

|      |      |      |      |      |
|------|------|------|------|------|
| 19.2 | 15.1 | 9.6  | 4.2  | 5.4  |
| 9.0  | 5.3  | 12.9 | 4.2  | 15.2 |
| 17.2 | 12.0 | 17.3 | 7.8  | 8.0  |
| 8.2  | 3.0  | 6.0  | 9.5  | 11.7 |
| 4.5  | 18.5 | 24.3 | 3.9  | 17.2 |
| 13.5 | 5.8  | 21.3 | 8.7  | 4.0  |
| 20.7 | 6.8  | 19.3 | 5.9  | 3.8  |
| 7.9  | 14.5 | 2.5  | 5.3  | 7.4  |
| 19.5 | 3.3  | 9.1  | 1.8  | 5.3  |
| 8.8  | 11.1 | 8.1  | 10.1 | 10.6 |
| 18.7 | 16.4 | 9.8  | 10.0 | 15.2 |
| 7.4  | 7.3  | 15.4 | 18.7 | 11.5 |
| 9.7  | 7.4  | 15.7 | 5.6  | 5.9  |
| 13.7 | 7.3  | 8.2  | 3.3  | 20.1 |
| 8.1  | 5.2  | 8.8  | 7.3  | 12.2 |
| 8.4  | 10.2 | 7.2  | 11.3 | 12.0 |
| 10.8 | 3.1  | 12.8 | 2.9  | 8.8  |

- (i) Using an appropriate probability model for the population from which the data is a random sample, obtain a *precise* 95% confidence interval for the mean of this population; use this interval estimate to test the hypothesis by the Editor-in-Chief that the mean time-to-publication is 9 months, against the alternative that it is higher.
- (ii) Considering  $n = 85$  as a large enough sample size for a normal approximation for the distribution of the sample mean, repeat (i) carrying out an appropriate one-sample test. Compare your result with that in (i). How good is the normal approximation?
- (iii) Use the normal approximation to test the hypothesis that the mean time to

publication is actually 10 months, versus the alternative that it is not. Interpret your result *vis à vis* the result in (ii).

**15.53** The data shown in the table below (see Problem 1.15) shows a four-year record of the number of “recordable” safety incidents occurring per month at a plant site.

|   |   |   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 1 | 0 | 1 |
| 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2 | 2 | 0 | 1 | 2 | 0 | 1 | 2 | 1 | 1 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |

- (i) If the company proposes to use as a “safety performance interval” (SPI), the statistic

$$S_I = \bar{X} \pm 3\sqrt{\bar{X}}$$

compute this interval from the supplied sample data.

(ii) The utility of the SPI is that any observation falling outside the *upper bound* is deemed to be indicative of a potential real increase in the number of safety incidents. Consider that over the most recent four-month period, the plant recorded 1, 3, 2, 3 safety incidents respectively. According to the SPI criterion, is there evidence that there has been a real increase in the number of incidents during any of the most recent four months?

**15.54** Refer to Problem 15.53 and consider the supplied data as a random sample from a Poisson population with unknown mean  $\lambda$ .

- (i) Assuming that a sample size of 48 is large enough for the normal approximation to be valid for the distribution of the sample mean, use the sample data to test the hypothesis that the population mean  $\lambda = 0.5$  versus the alternative that it is not.  
(ii) Use the *theoretical* value postulated for the population mean and compute the  $P(X \geq x_0 | \lambda = 0.5)$  for  $x_0 = 1, 2, 3$  and hence determine the *p*-values associated with the individual hypotheses that each recent observation, 1, 2, and 3, belongs to the same Poisson population,  $\mathcal{P}(0.5)$ , against the alternative that the observations belong to a different population with  $\lambda_a > 0.5$ .

**15.55** In clinics where “Assisted Reproductive Technologies” such as in-vitro fertilization are used to help infertile couples conceive and bear children (see Chapter 11 for a case study), it is especially important to be able to determine probability of a single embryo resulting in a live birth at the end of the treatment cycle. As shown in Chapter 11, determining this parameter, which is equivalent to a binomial probability of success, remains a challenge, however. A typical clinical study, the result of which may be used to determine this parameter for carefully selected cohort groups, is described below.

A cohort of 100 patients under the age of 35 years (the “Younger” group), and another cohort of the same size, but consisting of patients that are 35 years and older (the “Older” group), participated in a clinical study where each patient received five embryos in an in-vitro fertilization (IVF) treatment cycle. The results are shown in the table below:  $x$  is the number of live births per delivered pregnancy;  $y_O$  and  $y_Y$  represent, respectively, how many in the older and younger group had the pregnancy outcome of  $x$ .

- (i) At the  $\alpha = 0.05$  significance level, determine whether or not the single embryo probability of success parameter,  $p$ , is different for each cohort group.

- (ii) At the  $\alpha = 0.05$ , test the hypothesis that  $p = 0.3$  for the older cohort group versus the alternative that it is *less than* 0.3. Interpret your result.
- (iii) At the  $\alpha = 0.05$ , test the hypothesis that  $p = 0.3$  for the younger cohort group versus the alternative that it is *greater than* 0.3. Interpret your result.
- (iv) If it is desired to be able to determine the probability of success parameter for older cohort group to within  $\pm 0.05$  with 95% confidence, determine the size of the cohort group to use in the clinical study.

| $x$<br>No. of live<br>births in a<br>delivered<br>pregnancy | $y_O$<br>Total no. of<br>“older patients”<br>(out of 100)<br>with pregnancy outcome $x$ | $y_Y$<br>Total no. of<br>“younger patients”<br>(out of 100)<br>with pregnancy outcome $x$ |
|---|---|---|
| 0   | 32  | 8   |
| 1   | 41  | 25  |
| 2   | 21  | 35  |
| 3   | 5   | 23  |
| 4   | 1   | 8   |
| 5   | 0   | 1   |

**15.56** To characterize precisely how many sick days its employees take, a random sample of 50 employee files was selected and the following statistics were determined:  $\bar{x} = 9.50$  days and  $s^2 = 42.25$ .

- (i) Determine a 95% confidence interval on  $\mu$ , the true population mean number of sick days taken per employee.
- (ii) Does the evidence in the data support the hypothesis that the mean number of sick days taken by employees is less than 14.00 days?
- (iii) What is the power of the test you conducted in (ii). State any assumptions clearly. (iv) The personnel director who ordered the study was a bit surprised at how large the computed sample variance turned out to be. However, the human resources statistician insisted that this is not necessarily larger than the typical industry value of  $\sigma^2 = 35$ . Assuming that the sample is from a normal population, carry out an appropriate test to confirm or refute this claim. What is the  $p$ -value associated with the test?

**15.57** It is desired to characterize the precision of two instruments used to measure the density of a liquid stream in a refinery's distillation column. Ten measurements of a calibration sample with known density 0.85 gm/cc are shown in the table below.

| Instrument 1<br>Measurements | Instrument 2<br>Measurements |
|------------------------------|------------------------------|
| 0.864                        | 0.850                        |
| 0.858                        | 0.916                        |
| 0.855                        | 0.861                        |
| 0.764                        | 0.866                        |
| 0.791                        | 0.874                        |
| 0.827                        | 0.901                        |
| 0.849                        | 0.836                        |
| 0.818                        | 0.953                        |
| 0.747                        | 0.733                        |
| 0.846                        | 0.836                        |

Consider these data as random samples from two independent normal populations; carry out an appropriate test to confirm or refute the hypothesis that instrument 2 is less precise than instrument 1.

**15.58** In producing the enzyme cyclodextrin glucosyltransferase in bacterial cultures via two different methods (“shaken” and “surface”), Ismail *et al.*, (1996)<sup>10</sup>, obtained the data shown in the table below on the protein content (in mg/ml) obtained by each method.

| Protein content (mg/ml) |         |
|-------------------------|---------|
| Shaken                  | Surface |
| 1.91                    | 1.71    |
| 1.66                    | 1.57    |
| 2.64                    | 2.51    |
| 2.62                    | 2.30    |
| 2.57                    | 2.25    |
| 1.85                    | 1.15    |

Is the variability in the protein content the same for both methods? State any assumptions you may need to make in answering this question.

**15.59** The table below (see also Problem 14.41 in Chapter 14) shows the time in months between occurrences of safety violations for three operators, “A,” “B,” and “C,” working in a toll manufacturing facility.

|   |      |      |      |      |      |      |      |      |      |      |
|---|------|------|------|------|------|------|------|------|------|------|
| A | 1.31 | 0.15 | 3.02 | 3.17 | 4.84 | 0.71 | 0.70 | 1.41 | 2.68 | 0.68 |
| B | 1.94 | 3.21 | 2.91 | 1.66 | 1.51 | 0.30 | 0.05 | 1.62 | 6.75 | 1.29 |
| C | 0.79 | 1.22 | 0.65 | 3.90 | 0.18 | 0.57 | 7.26 | 0.43 | 0.96 | 3.76 |

Since the random variable in question is exponentially distributed and the sample size of 10 is considerably smaller than is required for a normal approximation to be valid for the sampling distribution of the sample mean, testing hypotheses about the difference between the means of these populations requires a different approach. The precise 95% confidence interval estimates of the unknown population parameters (obtained from the sampling distribution of the mean of an exponential random variable (Problem 14.41)) can be used to investigate if the population means overlap. An alternative approach involves the distribution of the difference between two exponential random variables.

It can be shown (Exercise 9.3) that given two independent random variables,  $X_1$  and  $X_2$ , with identical exponential  $\mathcal{E}(\beta)$  distributions, the pdf of their difference,

$$Y = X_1 - X_2 \quad (15.182)$$

is the double exponential (or Laplace) distribution defined as:

$$f(y) = \frac{1}{2\beta} e^{-|y|/\beta}; \quad -\infty < y < \infty \quad (15.183)$$

with mean 0 and variance  $2\beta^2$ . It can be shown that, in part because of the symmetric

---

<sup>10</sup>Ismail A.S, U.I. Sobieh, and A.F. Abdel-Fattah, (1996). “Biosynthesis of cyclodextrin glucosyltransferase and  $\beta$ -cyclodextrin by *Bacillus macerans* 314 and properties of the crude enzyme. *The Chem Eng. J.*, 61 247–253.

nature of this distribution, the distribution of  $\bar{Y}$ , the mean of a random sample of size  $n$  from this distribution, is approximately Gaussian with mean 0 and variance  $2\beta^2/n$ . More importantly, again, because of the symmetry of the distribution, the approximation is quite reasonable even for modest sample sizes as small as  $n = 10$ .

Form the differences  $Y_{AB} = X_A - X_B$  and  $Y_{AC} = X_A - X_C$  from the given data and use the normal approximation to the sampling distribution of the mean of a Laplace random variable to test the hypotheses that operator A is more safety conscious on the average than operator B, and also more than operator C. If you have to make any further assumptions, state them clearly.

**15.60** It is desired to determine if there is a significant difference in the average number of high-end vacuum cleaners produced per 8-hour workday by two different assembly plants located in two different countries. A random sample of 10 such daily outputs was selected for each assembly plant from last year's production tally, and the results summarized in the table below:

| Statistics         | Plant A | Plant B |
|--------------------|---------|---------|
| Sample Size, $n$   | 10      | 10      |
| Average, $\bar{x}$ | 28      | 33      |
| Variance, $s^2$    | 10.5    | 13.2    |

First ascertain whether or not the two population variances are the same. Then carry out an appropriate test of the equality of the mean production output that is commensurate with your findings regarding the variances. Interpret your results.

**15.61** In a certain metal oxide ore refining process, several samples (between 6 and 12) are taken monthly from various sections of the huge fluidized bed reactor and analyzed to determine average monthly residual silica content. The table below shows the result of such analyses for a 6-month period.

| Month | Number of Samples Analyzed | $\bar{x}$<br>Average Silica Content<br>(coded units) | $s$<br>Sample Standard Deviation |
|-------|----------------------------|--|----------------------------------|
| Jan   | 12                         | 65.8   | 32.1                             |
| Feb   | 9                          | 36.9   | 21.0                             |
| Mar   | 10                         | 49.4   | 24.6                             |
| Apr   | 11                         | 74.4   | 17.8                             |
| May   | 7                          | 59.3   | 15.2                             |
| Jun   | 11                         | 76.6   | 17.2                             |

The standard by which the ore refining operation is declared normal for any month requires a mean silica content  $\mu = 63.7$ , and inherent variability,  $\sigma = 21.0$ ; otherwise the operation is considered *abnormal*. At a 5% significance level, identify those months during which the refinery operation would be considered "abnormal." Support your conclusions adequately.

TABLE 15.12: Summary of Selected Hypothesis Tests and their Characteristics

| Population Parameter, $\theta$<br>(Null Hypothesis, $H_0$ )                  | Point Estimator, $\hat{\Theta}$   | Test Statistic  | Test                                     | $H_0$ Rejection Condition |
|--|---|---|--|---------------------------|
| $\mu; (H_0 : \mu = \mu_0)$<br>Small sample $n < 30$                          | $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$<br>( $S$ for unknown $\sigma$ )  | $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$<br>$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$   | $z$ -test<br>$t$ -test                   | Table 15.2<br>Table 15.3  |
| $\delta = \mu_1 - \mu_2; (H_0 : \delta = \delta_0)$<br>Small sample $n < 30$ | $\bar{D} = \bar{X}_1 - \bar{X}_2$<br>$(S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2})$                                  | $Z = \frac{\bar{D} - \delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$<br>$T = \frac{\bar{D} - \delta_0}{\sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$ | 2-sample $z$ -test<br>2-sample $t$ -test | Table 15.4<br>Table 15.5  |
| $\delta = \mu_1 - \mu_2; (H_0 : \delta = \delta_0)$<br>(Paired)              | $\bar{D} = \frac{\sum_{i=1}^n D_i}{n}$<br>$(D_i = X_{1_i} - X_{2_i})$<br>$(S_D^2 = \frac{\sum_{i=1}^n (D_i - \bar{D})^2}{n-1})$ | $T = \frac{\bar{D} - \delta_0}{S_D/\sqrt{n}}$   | Paired $t$ -test                         | Table 15.7                |
| $\sigma^2; (H_0 : \sigma^2 = \sigma_0^2)$                                    | $S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$  | $C^2 = \frac{(n-1)S^2}{\sigma_0^2}$   | Chi-squared-test                         | Table 15.9                |
| $\sigma_1^2/\sigma_2^2; (H_0 : \sigma_1^2 = \sigma_2^2)$                     | $S_1^2/S_2^2$   | $F = \frac{S_1^2}{S_2^2}$   | $F$ -test                                | Table 15.10               |



# **Chapter 16**

---

## **Regression Analysis**

|         |   |     |
|---------|---|-----|
| 16.1    | Introductory Concepts .....                       | 648 |
| 16.1.1  | Dependent and Independent Variables .....         | 648 |
| 16.1.2  | The Principle of Least Squares .....              | 651 |
| 16.2    | Simple Linear Regression .....                    | 652 |
| 16.2.1  | One-Parameter Model .....                         | 652 |
| 16.2.2  | Two-Parameter Model .....                         | 653 |
|         | Primary Model Assumption .....                    | 653 |
|         | Ordinary Least Squares (OLS) Estimates .....      | 654 |
|         | Maximum Likelihood Estimates .....                | 657 |
|         | Actual Regression Line and Residuals .....        | 657 |
| 16.2.3  | Properties of OLS Estimators .....                | 659 |
| 16.2.4  | Confidence Intervals .....                        | 661 |
|         | Slope and Intercept Parameters .....              | 661 |
|         | Regression Line .....                             | 663 |
| 16.2.5  | Hypothesis Testing .....                          | 664 |
| 16.2.6  | Prediction and Prediction Intervals .....         | 668 |
| 16.2.7  | Coefficient of Determination and the F-Test ..... | 670 |
|         | Orthogonal Decomposition of Variability .....     | 671 |
|         | $R^2$ , The Coefficient of Determination .....    | 672 |
|         | F-test for Significance of Regression .....       | 674 |
| 16.2.8  | Relation to the Correlation Coefficient .....     | 676 |
| 16.2.9  | Mean-Centered Model .....                         | 677 |
| 16.2.10 | Residual Analysis .....                           | 678 |
| 16.3    | “Intrinsically” Linear Regression .....           | 682 |
| 16.3.1  | Linearity in Regression Models .....              | 682 |
| 16.3.2  | Variable Transformations .....                    | 685 |
| 16.4    | Multiple Linear Regression .....                  | 686 |
| 16.4.1  | General Least Squares .....                       | 687 |
| 16.4.2  | Matrix Methods .....                              | 688 |
|         | Properties of the Estimates .....                 | 689 |
|         | Residuals Analysis .....                          | 691 |
| 16.4.3  | Some Important Special Cases .....                | 694 |
|         | Weighted Least Squares .....                      | 694 |
|         | Constrained Least Squares .....                   | 696 |
|         | Ridge Regression .....                            | 697 |
| 16.4.4  | Recursive Least Squares .....                     | 698 |
|         | Problem Formulation .....                         | 698 |
|         | Recursive Least Squares Estimation .....          | 699 |
| 16.5    | Polynomial Regression .....                       | 700 |
| 16.5.1  | General Considerations .....                      | 700 |
| 16.5.2  | Orthogonal Polynomial Regression .....            | 704 |
|         | An Example: Gram Polynomials .....                | 704 |
|         | Application in Regression .....                   | 708 |
| 16.6    | Summary and Conclusions .....                     | 710 |
|         | REVIEW QUESTIONS .....                            | 711 |

|                            |     |
|----------------------------|-----|
| EXERCISES .....            | 713 |
| APPLICATION PROBLEMS ..... | 719 |

*The mathematical facts worthy of being studied  
 are those which, by their analogy with other facts  
 are capable of leading us to the knowledge of a mathematical law  
 just as experimental facts lead us  
 to the knowledge of a physical law ...*

Henri Poicaré (1854–1912)

It is often the case in many practical problems that the variability observed in a random variable,  $Y$ , consists of more than just the purely randomly varying phenomena that have occupied our attention up till now. For this new class of problems, an underlying functional relationship exists between  $Y$  and an independent variable,  $x$ , (deliberately written in the lower case for reasons that will soon become clear), with a purely random component superimposed on this otherwise deterministic component. This chapter is devoted to dealing with problems of this kind. The values observed for the random variable  $Y$  depend on the values of the (deterministic) variable,  $x$ , and, were it not for the presence of the purely random component,  $Y$  would have been perfectly predictable given  $x$ . Regression analysis is concerned with obtaining, from data, the best estimate of the relationship between  $Y$  and  $x$ .

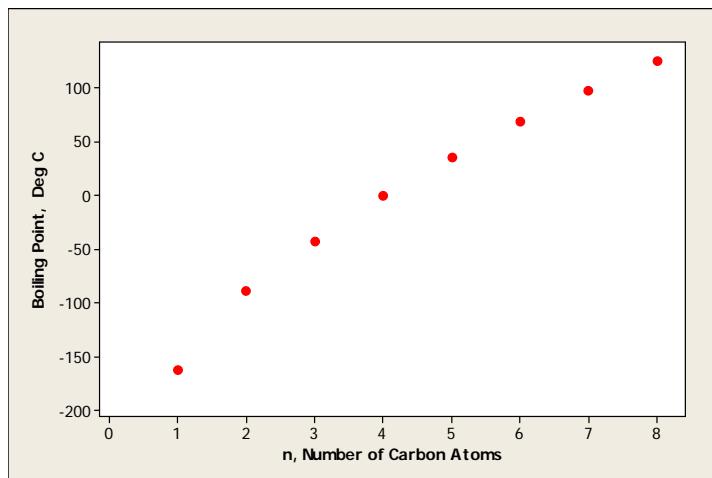
Although apparently different from what we have dealt with up until now, we will see that regression analysis in fact builds directly upon many of the results obtained thus far, especially estimation and hypothesis testing.

## 16.1 Introductory Concepts

Consider the data in Table 16.1 showing the boiling point (in  $^{\circ}\text{C}$ ) of 8 hydrocarbons in a homologous series, along with  $n$ , the number of carbon atoms in each molecule. A scatter plot of boiling point versus  $n$  is shown in Fig 16.1, where we notice right away that as the number of carbon atoms in this homologous series increases, so does the boiling point of the hydrocarbon compound. In fact, the implied relationship between these two variables appears to be so strong that one is immediately inclined to conclude that it must be possible to predict the boiling point of compounds in this series on the basis of the number of carbon atoms. There is therefore no doubt that there is some sort of a functional relationship between  $n$  and boiling point. If determined “correctly,” such a relationship will provide, among other things, a simple way to capture the extensive data on such “physical properties” of compounds in this particular homologous series.

**TABLE 16.1:** Boiling points of a series of hydrocarbons

| Hydrocarbon Compound | $n$ , Number of Carbon Atoms | Boiling Point $^{\circ}\text{C}$ |
|----------------------|------------------------------|----------------------------------|
| Methane              | 1                            | -162                             |
| Ethane               | 2                            | -88                              |
| Propane              | 3                            | -42                              |
| n-Butane             | 4                            | 1                                |
| n-Pentane            | 5                            | 36                               |
| n-Hexane             | 6                            | 69                               |
| n-Heptane            | 7                            | 98                               |
| n-Octane             | 8                            | 126                              |



**FIGURE 16.1:** Boiling point of hydrocarbons in Table 16.1 as a function of the number of carbon atoms in the compound

### 16.1.1 Dependent and Independent Variables

Many cases such as the one illustrated above arise in science and engineering where the value taken by one variable appears to depend on the value taken by another. Not surprisingly, it is customary to refer the variable whose value depends on the value of another as the *dependent* variable, while the other variable is known as the *independent* variable. It is often desired to capture the relationship between these two variables in some mathematical form. However, because of measurement errors and other sources of variability, this exercise requires the use of probabilistic and statistical techniques. Under these circumstances, the independent variable is considered as a fixed, deterministic quantity that is not subject to random variability. This is perfectly exemplified in  $n$ , the number of carbon atoms in the hydrocarbon compounds of Table 16.1; it is a known quantity not subject to random variability. The dependent variable, on the other hand, is the random variable, subject to a wide variety of potential sources of random variability, including, but not limited to measurement uncertainties. The dependent variable is therefore represented as the random variable,  $Y$ , while the independent variable is represented as the deterministic variable,  $x$ , represented in the lower case to underscore its deterministic nature.

The variability observed in the random variable,  $Y$ , is typically considered to consist of two distinct components, i.e., for each observation,  $Y_i, i = 1, 2, \dots, n$ :

$$Y_i = g(x_i; \boldsymbol{\theta}) + \epsilon_i \quad (16.1)$$

where  $g(x_i; \boldsymbol{\theta})$  is the deterministic component, a functional relationship, with  $\boldsymbol{\theta}$  as a set of unknown parameters, and  $\epsilon_i$  is the random component. The deterministic mathematical relationship between these two variables is a “model” of how the independent  $x$  (also known as the “predictor”) affects the predictable part of the dependent  $Y$ , sometimes known as the “response.”

In some cases, the functional form of  $g(x_i)$  is known from fundamental scientific principles. For example, if  $Y$  is the distance (in cms) traveled in time  $t_i$  secs by a particle launched with an initial velocity,  $u$  (cm/sec), and traveling at a constant acceleration  $a$  (cm/sec<sup>2</sup>), then we know that

$$g(t_i; u, a) = ut_i + \frac{1}{2}at_i^2 \quad (16.2)$$

with  $\boldsymbol{\theta} = (u, a)$  as the parameters.

In most cases, however, there is no such fundamental scientific principle to suggest an appropriate form for  $g(x_i; \boldsymbol{\theta})$ ; simple forms (typically polynomials) are postulated and validated with data, as we show subsequently. The result in this case is known as an “empirical” model because it is strictly dependent on data and not on some known fundamental scientific principle.

Regression analysis to be primarily concerned with the following tasks:

- Obtaining the “best estimates”  $\hat{\boldsymbol{\theta}}$  for the model parameters,  $\boldsymbol{\theta}$ ;

- Characterizing the random sequence  $\epsilon_i$ ; and,
- Making inference about the parameter estimates,  $\hat{\theta}$ .

The classical treatment is based on “least squares estimation” which we will discuss briefly now, before using it in the context of regression.

### 16.1.2 The Principle of Least Squares

Consider the case where the random sample,  $Y_1, Y_2, \dots, Y_n$ , is drawn from a population characterized by a single, constant parameter,  $\theta$ , the population mean. The random variable  $Y$  may then be written as:

$$Y_i = \theta + \epsilon_i \quad (16.3)$$

where the observed random variability is due to random component  $\epsilon_i$ . Furthermore, let the variance of  $Y$  be  $\sigma^2$ . Then from Eq (16.3), we obtain:

$$E[Y_i] = \theta + E[\epsilon_i] \quad (16.4)$$

and since, by definition,  $E[Y_i] = \theta$ , this implies that  $E[\epsilon_i] = 0$ . Furthermore,

$$\text{Var}(Y_i) = \text{Var}(\epsilon_i) = \sigma^2 \quad (16.5)$$

since  $\theta$  is a constant. Thus, from the fact that  $Y$  has a distribution (unspecified) with mean  $\theta$  and variance  $\sigma^2$  implies that in Eq (16.3), the random “error” term,  $\epsilon_i$  has zero mean and variance  $\sigma^2$ .

To estimate  $\theta$  from the given random sample, it seems reasonable to choose a value that is “as close as possible” to all the observed data. This concept may be represented mathematically as:

$$\min_{\theta} S(\theta) = \sum_{i=1}^n (Y_i - \theta)^2 \quad (16.6)$$

The usual calculus approach to this optimization problem leads to:

$$\left. \frac{\partial S}{\partial \theta} \right|_{\theta=\hat{\theta}} = -2 \sum_{i=1}^n (Y_i - \hat{\theta}) = 0 \quad (16.7)$$

which, when solved, produces the result:

$$\hat{\theta} = \frac{\sum_{i=1}^n Y_i}{n} \quad (16.8)$$

A second derivative with respect to  $\theta$  yields

$$\frac{\partial^2 S}{\partial \theta^2} = 2n > 0 \quad (16.9)$$

so that indeed  $S(\theta)$  achieves a minimum for  $\theta = \hat{\theta}$  in Eq (20.3).

The quantity  $\hat{\theta}$  in Eq (20.3) is referred to as a least-squares estimator for  $\theta$  in Eq (16.3), for the obvious reason that the value produced by this estimator achieves the minimum for the sum-of-squared deviation implied in Eq (16.6). It should not be lost on the reader that this estimator is also precisely the same as the familiar sample average.

The problems we have dealt with up until now may be represented in the form shown in Eq (16.3). In that context, the probability models we developed earlier may now be interpreted as models for  $\epsilon_i$ , the random variation around the constant random variable mean. This allows us to put the upcoming discussion on the regression problem in context of the earlier discussions.

Finally, we note that the principle of least-squares also affords us the flexibility to treat each observation,  $Y_i$ , differently in how it contributes to the estimation of  $\theta$ . This is done by applying appropriate weights  $W_i$  to Eq (16.3) to obtain:

$$W_i Y_i = W_i \theta + W_i \epsilon_i \quad (16.10)$$

Consequently, for example, more reliable observations can be assigned larger weights than less reliable ones. Upon using the same calculus techniques, the least-squares estimate in this case can be shown to be (see Exercise 16.2) to be:

$$\hat{\theta}_\omega = \frac{\sum_{i=1}^n W_i^2 Y_i}{\sum_{i=1}^n W_i^2} = \sum_{i=1}^n \omega_i Y_i \quad (16.11)$$

where

$$\omega_i = \frac{W_i^2}{\sum_{i=1}^n W_i^2} \quad (16.12)$$

Note that  $0 < \omega_i < 1$ . The result in Eq 16.11 is therefore an appropriately weighted average — a generalization of Eq (20.3) where  $\omega_i = 1/n$ . This variation on the least-squares approach is known appropriately as “weighted least-squares;” we shall encounter it later in this chapter.

## 16.2 Simple Linear Regression

### 16.2.1 One-Parameter Model

As a direct extension of Eq (16.3), let the relationship between the random variable  $Y$  and the independent (deterministic) variable,  $x$ , be:

$$Y = \theta x + \epsilon \quad (16.13)$$

where the random error,  $\epsilon$ , has zero mean and constant variance,  $\sigma^2$ . Then,  $E(Y|x)$ , the conditional expectation of  $Y$  given a specific value for  $x$  is:

$$\mu_{Y|x} = E(Y|x) = \theta x, \quad (16.14)$$

recognizable as the equation of a straight line with slope  $\theta$  and zero intercept. It is also known as the “one-parameter” regression model, a classic example of which is the famous Ohm’s law in physics: the relationship between the Voltage,  $V$ , across a resistor with unknown resistance,  $R$ , and the current  $I$  flowing through the resistive element, i.e.,

$$V = IR \quad (16.15)$$

From data  $y_i; i = 1, 2, \dots, n$ , actual values of the random variable,  $Y_i$ , observed for corresponding values of  $x_i$ , the problem at hand is to obtain an estimate of the characterizing parameter  $\theta$ . Using the method of least-squares outlined above requires minimizing the sum-of-squares function:

$$S(\theta) = \sum_{i=1}^n (y_i - \theta x_i)^2 \quad (16.16)$$

from where  $\partial S / \partial \theta = 0$  yields:

$$-2 \sum_{i=1}^n x_i (y_i - \theta x_i) = 0 \quad (16.17)$$

which, is solved for  $\theta$  to obtain:

$$\hat{\theta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \quad (16.18)$$

This is the expression for the slope of the “best” (i.e., least-squares) straight line (with zero intercept) through the points  $(x_i, y_i)$ .

### 16.2.2 Two-Parameter Model

More general is the two-parameter model,

$$Y = \theta_0 + \theta_1 x + \epsilon \quad (16.19)$$

indicating a functional relationship,  $g(x; \boldsymbol{\theta})$ , that is a straight line with slope  $\theta_1$  and potentially non-zero intercept  $\theta_0$  as the parameters, i.e.,

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} \quad (16.20)$$

along with  $E(\epsilon) = 0$ ;  $Var(\epsilon) = \sigma^2$ . In this case, the conditional expectation of  $Y$  given a specific value for  $x$  is given by:

$$\mu_{Y|x} = E(Y|x) = \theta_0 + \theta_1 x \quad (16.21)$$

In this particular case, regression analysis is primarily concerned with obtaining the best estimates for  $\hat{\boldsymbol{\theta}} = (\hat{\theta}_0, \hat{\theta}_1)$ ; characterizing the random sequence  $\epsilon_i$ ; and, making inference about the parameter estimates,  $\hat{\boldsymbol{\theta}} = (\hat{\theta}_0, \hat{\theta}_1)$ .

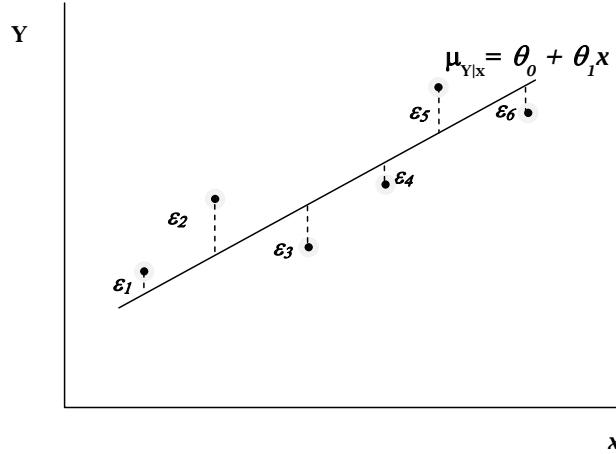


FIGURE 16.2: The true regression line and the zero mean random error  $\epsilon_i$

### Primary Model Assumption

In this case, the true but unknown regression line is represented by Eq (16.21), with data scattered around it. The fact that  $E(\epsilon) = 0$ , indicates that the data scatters “evenly” around the true line; more precisely, the data varies randomly around a mean value that is the function of  $x$  defined by the true but unknown regression line in Eq (16.21). This is illustrated in Figure 16.2

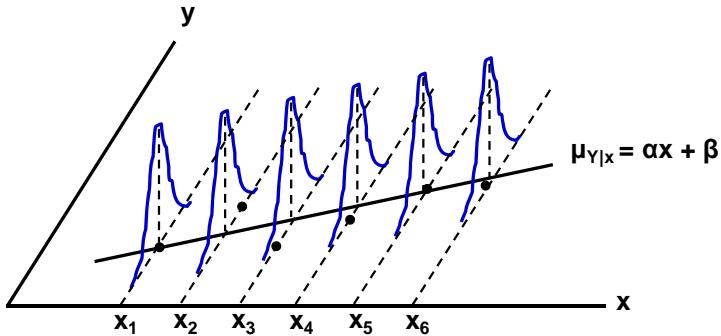
It is typical to assume that each  $\epsilon_i$ , the random component of the model, is mutually independent of the others and follows a Gaussian distribution with zero mean and variance  $\sigma^2$ , i.e.,  $\epsilon_i \sim N(0, \sigma^2)$ . The implication in this particular case is therefore that each data point,  $(x_i, y_i)$ , comes from a Gaussian distribution whose mean is dependent on the value of  $x$ , and falls on the true regression line, as illustrated in Fig 16.3. Equivalently, the true regression line passes through the mean of the series of Gaussian distributions having the same variance. The two main assumptions underlying regression analysis may now be summarized as follows:

1.  $\epsilon_i$  forms an independent random sequence, with zero mean and variance  $\sigma^2$  that is constant for all  $x$ ;
2.  $\epsilon_i \sim N(0, \sigma^2)$  so that  $Y_i \sim (\theta_0 + \theta_1 x, \sigma^2)$

### Ordinary Least Squares (OLS) Estimates

Obtaining the least-squares estimates of the intercept,  $\theta_0$ , and slope,  $\theta_1$ , from data  $(x_i, y_i)$  involves minimizing the sum-of-squares function,

$$S(\theta_0, \theta_1) = \sum_{i=1}^n [y_i - (\theta_1 x_i + \theta_0)]^2 \quad (16.22)$$



**FIGURE 16.3:** The Gaussian assumption regarding variability around the true regression line giving rise to  $\epsilon \sim N(0, \sigma^2)$ : The 6 points represent the data at  $x_1, x_2, \dots, x_6$ ; the solid straight line is the true regression line which passes through the mean of the sequence of the indicated Gaussian distributions

where the usual first derivatives of the calculus approach yield:

$$\frac{\partial S}{\partial \theta_0} = 2 \sum_{i=1}^n [y_i - (\theta_1 x_i + \theta_0)] = 0 \quad (16.23)$$

$$\frac{\partial S}{\partial \theta_1} = -2 \sum_{i=1}^n x_i [y_i - (\theta_1 x_i + \theta_0)] = 0 \quad (16.24)$$

These expressions rearrange to give:

$$\theta_1 \sum_{i=1}^n x_i + \theta_0 n = \sum_{i=1}^n y_i \quad (16.25)$$

$$\theta_1 \sum_{i=1}^n x_i^2 + \theta_0 \sum_{i=1}^n x_i = \sum_{i=1}^n x_i y_i \quad (16.26)$$

collectively known as the “normal equations,” to be solved simultaneously to produce the least squares estimates,  $\hat{\theta}_0$ , and  $\hat{\theta}_1$ .

Before solving these equations explicitly, we wish to direct the reader’s attention to a pattern underlying the emergence of the normal equations. Beginning with the original two-parameter model equation:

$$y_i = \theta_1 x_i + \theta_0 + \epsilon_i$$

a summation across each term yields:

$$\sum_{i=1}^n y_i = \theta_1 \sum_{i=1}^n x_i + \theta_0 n \quad (16.27)$$

where the last term involving  $\epsilon_i$  has vanished upon the assumption that  $n$  is

sufficiently large so that because  $E(\epsilon_i) = 0$ , the sum will be close to zero (a point worth keeping in mind to remind the reader that the result of solving the normal equations provide estimates, not “precise” values).

Also, multiplying the model equation by  $x_i$  and summing yields:

$$\sum_{i=1}^n y_i x_i = \theta_1 \sum_{i=1}^n x_i^2 + \theta_0 \sum_{i=1}^n x_i \quad (16.28)$$

where, once again the last term involving  $\epsilon_i$  has vanished because of independence with  $x_i$  and the assumption, once again, that  $n$  is sufficiently large that the sum will be close to zero. Note that these two equations are identical to the normal equations; more importantly, as derived by summation from the original model they are the *sample equivalents* of the following expectations:

$$E(Y) = \theta_1 E(x) + \theta_0 \quad (16.29)$$

$$E(Yx) = \theta_1 E(x^2) + \theta_0 E(x) \quad (16.30)$$

which should help put the emergence of the normal equations into perspective.

Returning to the task of computing least squares estimates of the two model parameters, let us define the following terms:

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 \quad (16.31)$$

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (16.32)$$

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad (16.33)$$

where  $\bar{y} = (\sum_{i=1}^n y_i)/n$  and  $\bar{x} = (\sum_{i=1}^n x_i)/n$  represent the usual averages. When expanded out and consolidated, these equations yield:

$$nS_{xx} = n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2 \quad (16.34)$$

$$nS_{yy} = n \sum_{i=1}^n y_i^2 - \left( \sum_{i=1}^n y_i \right)^2 \quad (16.35)$$

$$nS_{xy} = n \sum_{i=1}^n x_i y_i - \left( \sum_{i=1}^n x_i \right) \left( \sum_{i=1}^n y_i \right) \quad (16.36)$$

These terms, clearly related to sample variances and covariances, allow us to solve Eqns (16.25) and (16.26) simultaneously to obtain the results:

$$\hat{\theta}_1 = \frac{S_{xy}}{S_{xx}} \quad (16.37)$$

$$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x} \quad (16.38)$$

Nowadays, such computations implied in this derivation are no longer carried out by hand, of course, but by computer programs; the foregoing discussion is therefore intended to acquaint the reader with the principles and mechanics underlying the numbers produced by the statistical software packages.

### Maximum Likelihood Estimates

Under the Gaussian assumption, the regression equation, written in the more general form,

$$Y = \eta(x, \boldsymbol{\theta}) + \epsilon, \quad (16.39)$$

implies that the observations  $Y_1, Y_2, \dots, Y_n$  come from a Gaussian distribution with mean  $\eta$  and variance,  $\sigma^2$ ; i.e.  $Y \sim N(\eta(x, \boldsymbol{\theta}), \sigma^2)$ . If the data can be considered as a random sample from this distribution, then the method of maximum likelihood presented in Chapter 14 may be used to estimate  $\eta(x, \boldsymbol{\theta})$  and  $\sigma^2$  in precisely the same manner in which estimates of the  $N(\mu, \sigma^2)$  population parameters were determined in Section 14.3.2. The only difference this time is that the population mean,  $\eta(x, \boldsymbol{\theta})$ , is no longer constant, but a function of  $x$ . It can be shown (see Exercise 16.5) that when the variance  $\sigma^2$  is constant, the maximum likelihood estimate for  $\theta$  in the one-parameter model,

$$\eta(x, \boldsymbol{\theta}) = \theta x \quad (16.40)$$

and the maximum likelihood estimates for  $(\theta_0, \theta_1)$  in the two-parameter model,

$$\eta(x, \boldsymbol{\theta}) = \theta_0 + \theta_1 x \quad (16.41)$$

are each identical to the corresponding least squares estimates obtained in Eq (19.49) and in Eqs (16.38) and (16.37) respectively. It can also be shown (see Exercise 16.6) that when the variance,  $\sigma_i^2$ , associated with each observation,  $Y_i$ ,  $i = 1, 2, \dots, n$ , differs from observation to observation, the maximum likelihood estimates for the parameters  $\theta$  in the first case, and for  $(\theta_0, \theta_1)$  in the second case, are the same as the corresponding weighted least squares estimates, with weights related to the reciprocal of  $\sigma_i$ .

### Actual Regression Line and Residuals

In the same manner in which the true (constant) mean,  $\mu$ , of a Gaussian distribution producing the random sample  $X_1, X_2, \dots, X_n$ , is not known, only estimated by the sample average  $\bar{X}$ , the true regression line is also never known but estimated. When the least-squares estimates  $\hat{\theta}_0$  and  $\hat{\theta}_1$  are introduced into the original model, the result is the estimated observation  $\hat{y}$  defined by:

$$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x \quad (16.42)$$

This is not the same as the true theoretical  $\mu_{Y|x}$  in Eq (16.21) because, in general  $\hat{\theta}_0 \neq \theta_0$  and  $\hat{\theta}_1 \neq \theta_1$ ;  $\hat{y}_i$  is the two-parameter model's best estimate

**TABLE 16.2:**  
Density (in gm/cc) and  
weight percent of ethanol  
in ethanol-water mixture

| Density<br>(g/cc) | Wt %<br>Ethanol |
|-------------------|-----------------|
| 0.99823           | 0               |
| 0.98938           | 5               |
| 0.98187           | 10              |
| 0.97514           | 15              |
| 0.96864           | 20              |
| 0.96168           | 25              |
| 0.95382           | 30              |
| 0.94494           | 35              |

(or prediction) of the true but unknown value of the observation  $y_i$  (unknown because of the additional random effect,  $\epsilon_i$ ). If we now define as  $e_i$ , the error between the actual observation and the estimated value, i.e.,

$$e_i = y_i - \hat{y}_i, \quad (16.43)$$

this term is known as the residual error or simply the “residual;” it is our best estimate of the unknown  $\epsilon_i$ , just as  $\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x$  is our best estimate of the true regression line  $\mu_{Y|x} = E(Y|x) = \theta_1 x + \theta_0$ .

As discussed shortly (Section 16.2.10), the nature of the sequence of residuals provides a great deal of information about how well the model represents the observations.

**Example 16.1: DENSITY OF ETHANOL-WATER MIXTURE**

An experimental investigation into how the density of an ethanol-water mixture varies with weight percent of ethanol in the mixture yielded the result shown in Table 16.2. Postulate a linear two-parameter model as in Eq (16.19), and use the supplied data to obtain least-squares estimates of the slope and intercept, and also the residuals. Plot the data versus the model and comment on the fit.

**Solution:**

Given this data set, just about any software package, from Excel to MATLAB and MINITAB, will produce the following estimates:

$$\hat{\theta}_1 = -0.001471; \hat{\theta}_0 = 0.9975 \quad (16.44)$$

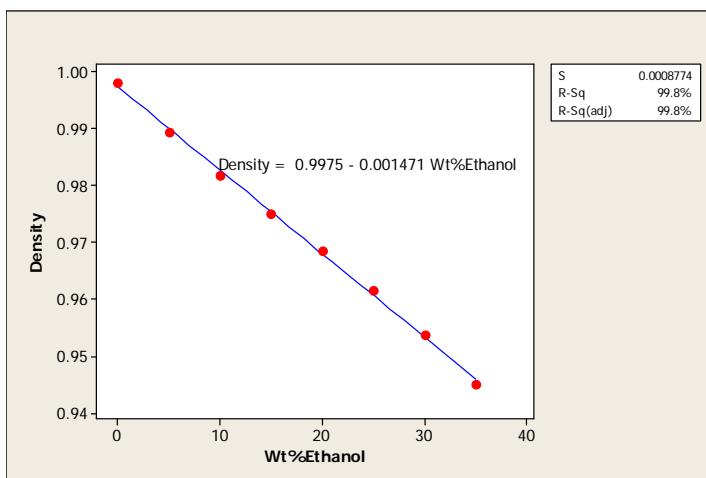
so that, if  $y$  is the density and  $x$  is the wt % of ethanol, the regression model fit to this data is given as:

$$\hat{y} = -0.001471x + 0.9975 \quad (16.45)$$

The model fit to the data is shown in Fig 16.4; and for the given values

**TABLE 16.3:** Density and weight percent of ethanol in ethanol-water mixture: model fit and residual errors

| Density(g/cc)<br><i>y</i> | Wt % Ethanol<br><i>x</i> | Estimated<br>Density, $\hat{y}$ | Residual<br>Errors, <i>e</i> |
|---------------------------|--------------------------|---------------------------------|------------------------------|
| 0.99823                   | 0                        | 0.997500                        | 0.000730                     |
| 0.98938                   | 5                        | 0.990145                        | -0.000765                    |
| 0.98187                   | 10                       | 0.982790                        | -0.000920                    |
| 0.97514                   | 15                       | 0.975435                        | -0.000295                    |
| 0.96864                   | 20                       | 0.968080                        | 0.000560                     |
| 0.96168                   | 25                       | 0.960725                        | 0.000955                     |
| 0.95382                   | 30                       | 0.953370                        | 0.000450                     |
| 0.94494                   | 35                       | 0.946015                        | -0.001075                    |



**FIGURE 16.4:** The fitted straight line to the Density versus Ethanol Weight % data: The additional terms included in the graph, *S*, *R-Sq* and *R-Sq(adj)* are discussed later

of *x*, the estimated  $\hat{y}$ , and the residuals, *e*, are shown in Table 16.3. Visually, the model seems to fit quite well.

This model allows us to predict solution density for any given weight percent of ethanol within the experimental data range but not actually part of the data. For example, for  $x = 7.5$ , Eq (16.45) estimates  $\hat{y} = 0.98647$ . How the residuals are analyzed is discussed in Section 16.2.10.

Expressions such as the one obtained in this example, Eq (16.45), are sometimes known as calibration curves. Such curves are used to calibrate measurement devices such as thermocouples, where the raw instrument output (say millivolts) is converted to the actual desired measurement (say temperature in  $^{\circ}\text{C}$ ) based on expressions such as the one obtained here. Such expressions are typically generated from standardized experiments where data on instrument output are gathered for various objects with known temperature.

### 16.2.3 Properties of OLS Estimators

When experiments are repeated for the same fixed values  $x_i$ , as a typical consequence of random variation, the corresponding value observed for  $Y_i$  will differ each time. The resulting estimates provided in (16.38) and Eqs (16.37) therefore will also change slightly each time. In typical fashion, therefore, the specific parameter estimates  $\hat{\theta}_0$  and  $\hat{\theta}_1$  are properly considered as realizations of the respective estimators  $\Theta_0$  and  $\Theta_1$ , random variables that depend on the random sample  $Y_1, Y_2, \dots, Y_n$ . It will be desirable to investigate the theoretical properties of these estimators defined by:

$$\Theta_1 = \frac{S_{xy}}{S_{xx}} \quad (16.46)$$

$$\Theta_0 = \bar{Y} - \Theta_1 \bar{x} \quad (16.47)$$

Let us begin with the expected values of these estimators. From here, we observe that

$$E(\Theta_1) = E\left(\frac{S_{xy}}{S_{xx}}\right) \quad (16.48)$$

which, from the definitions given above, becomes:

$$E(\Theta_1) = \frac{1}{S_{xx}} E\left[\sum_{i=1}^n Y_i(x_i - \bar{x})\right] \quad (16.49)$$

(because  $\sum_{i=1}^n \bar{Y}(x_i - \bar{x}) = 0$ , since  $\bar{Y}$  is a constant); and upon introducing Eq (16.19) in for  $Y_i$ , we obtain:

$$E(\Theta_1) = \frac{1}{S_{xx}} E\left[\sum_{i=1}^n (\theta_1 x_i + \theta_0 + \epsilon_i)(x_i - \bar{x})\right] \quad (16.50)$$

A term-by-term expansion and subsequent simplification results in

$$E(\Theta_1) = \frac{1}{S_{xx}} E\left[\theta_1 \sum_{i=1}^n (x_i - \bar{x})\right] \quad (16.51)$$

because  $\sum_{i=1}^n (x_i - \bar{x}) = 0$  and  $E[\sum_{i=1}^n \epsilon_i(x_i - \bar{x})] = 0$  since  $E(\epsilon_i) = 0$ . Hence, Eq (16.51) simplifies to

$$E(\Theta_1) = \frac{1}{S_{xx}} \theta_1 S_{xx} = \theta_1 \quad (16.52)$$

indicating that  $\Theta_1$  is an unbiased estimator of  $\theta_1$ , the true slope.

Similarly, from Eq (16.47), we obtain:

$$E(\Theta_0) = E(\bar{Y} - \Theta_1 \bar{x}) = E(\bar{Y}) - E(\Theta_1) \bar{x} \quad (16.53)$$

which by virtue of Eq (16.51) simplifies to:

$$E(\Theta_0) = \theta_1 \bar{x} + \theta_0 - \theta_1 \bar{x} = \theta_0 \quad (16.54)$$

so that  $\Theta_0$  is also an unbiased estimator for  $\theta_0$ , the true intercept.

In similar fashion, by definition of the variance of a random variable, it is straightforward to show that:

$$Var(\Theta_1) = \sigma_{\Theta_1}^2 = \frac{\sigma^2}{S_{xx}} \quad (16.55)$$

$$Var(\Theta_0) = \sigma_{\Theta_0}^2 = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \quad (16.56)$$

where  $\sigma^2$  is the variance of the random component,  $\epsilon$ . Consequently, the standard error of each estimate, the positive square root of the variance, is given by:

$$SE(\Theta_1) = \frac{\sigma}{\sqrt{S_{xx}}} \quad (16.57)$$

$$SE(\Theta_0) = \sigma \sqrt{\left( \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)} \quad (16.58)$$

#### 16.2.4 Confidence Intervals

As with all estimation problems, the point estimates obtained above for the regression parameters,  $\theta_0$  and  $\theta_1$ , by themselves are insufficient in making decisions about their true, but unknown values; we must add a measure of how precise these estimates are. Obtaining interval estimates is one option; and such interval estimates are determined for regression parameters essentially by the same procedure as that presented in Chapter 14 for population parameters. This, of course, requires sampling distributions.

##### Slope and Intercept Parameters

Under the Gaussian distributional assumption for  $\epsilon$ , with the implication that the sample  $Y_1, Y_2, \dots, Y_n$ , possesses the distribution  $N(\theta_0 + \theta_1 x, \sigma^2)$ , and from the results obtained above about the characteristics of the estimates, it can be shown that the random variables  $\Theta_1$  and  $\Theta_0$ , respectively the slope and the intercept, are distributed as  $\Theta_1 \sim N(\theta_1, \sigma_{\Theta_1}^2)$  and  $\Theta_0 \sim N(\theta_0, \sigma_{\Theta_0}^2)$  with the variances as shown in Eqns (16.55) and (16.56), *provided the data variance,  $\sigma^2$ , is known*. However, this variance is not known and must be estimated from data. This is done as follows for this particular problem.

Consider residual errors,  $e_i$ , our best estimates of  $\epsilon_i$ ; define the residual

error sum of squares as

$$SS_E = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (16.59)$$

$$\begin{aligned} &= \sum_{i=1}^n [y_i - (\hat{\theta}_1 x_i + \hat{\theta}_0)]^2 \\ &= \sum_{i=1}^n [(y_i - \bar{y}) - \hat{\theta}_1(x_i - \bar{x})]^2 \end{aligned} \quad (16.60)$$

which, upon expansion and simplification reduces to:

$$SS_E = S_{yy} - \hat{\theta}_1 S_{xy} \quad (16.61)$$

It can be shown that

$$E(SS_E) = (n - 2)\sigma^2 \quad (16.62)$$

as a result, the mean squared error,  $s_e^2$ , defined as:

$$s_e^2 = \frac{SS_E}{(n - 2)} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2} \quad (16.63)$$

is an unbiased estimate of  $\sigma^2$ .

Now, as with previous statistical inference problems concerning normal populations with unknown  $\sigma$ , by substituting  $s_e^2$ , the mean residual sum-of-squares, for  $\sigma^2$ , we have the following results: the statistics  $T_1$  and  $T_0$  defined as:

$$T_1 = \frac{\Theta_1 - \theta_1}{s_e / \sqrt{S_{xx}}} \quad (16.64)$$

and

$$T_0 = \frac{\Theta_0 - \theta_0}{s_e \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \quad (16.65)$$

each possess  $t$ -distribution with  $\nu = n - 2$  degrees of freedom. The immediate implications are therefore that

$$\theta_1 = \hat{\theta}_1 \pm t_{\alpha/2}(n - 2) \frac{s_e}{\sqrt{S_{xx}}} \quad (16.66)$$

$$\theta_0 = \hat{\theta}_0 \pm t_{\alpha/2}(n - 2) s_e \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)} \quad (16.67)$$

constitute  $(1 - \alpha) \times 100\%$  confidence intervals around the slope and intercept estimates, respectively.

**Example 16.2: CONFIDENCE INTERVAL ESTIMATES FOR THE SLOPE AND INTERCEPT OF ETHANOL-WATER MIXTURE DENSITY REGRESSION MODEL**

Obtain 95% confidence interval estimates for the slope and intercept of the regression model obtained in Example 16.1 for the ethanol-water mixture density data.

**Solution:**

In carrying out the regression in Example 16.1 with MINITAB, part of the computer program output is the set of standard errors. In this case,  $SE(\Theta_1) = 0.00002708$  for the slope, and  $SE(\Theta_0) = 0.000566$  for the intercept. These could also be computed by hand (although not recommended). Since the data set consists of 8 data points, we obtain the required  $t_{0.025}(6) = 2.447$  from the cumulative probability feature. The required 95% confidence intervals are therefore obtained as follows:

$$\theta_1 = -0.001471 \pm 0.00006607 \quad (16.68)$$

$$\theta_0 = 0.9975 \pm 0.001385 \quad (16.69)$$

Note that none of these two intervals includes 0.

### Regression Line

The actual regression line fit (see for example Fig 16.4), an estimate of the true but unknown regression line, is obtained by introducing into Eq (16.21), the estimates for the slope and intercept parameters to give

$$\hat{\mu}_{Y|x} = \hat{\theta}_1 x + \hat{\theta}_0 \quad (16.70)$$

For any specific value  $x = x^*$ , the value

$$\hat{\mu}_{Y|x^*} = \hat{\theta}_1 x^* + \hat{\theta}_0 \quad (16.71)$$

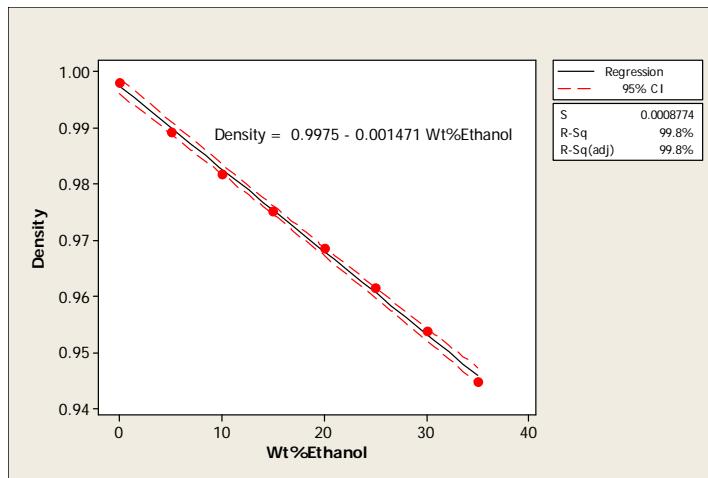
is the estimate of the actual response of  $Y$  at this point (akin to the sample average estimate of a true but unknown population mean).

In the same manner in which we obtained confidence intervals for sample averages, we can also obtain a confidence interval for  $\hat{\mu}_{Y|x^*}$ . It can be shown from Eq (16.71) (and Eq (16.56)) that the associated variance is:

$$Var(\hat{\mu}_{Y|x^*}) = \sigma^2 \left[ \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right] \quad (16.72)$$

and because of the normality of the random variables  $\Theta_0$  and  $\Theta_1$ , then if  $\sigma$  is known,  $\hat{\mu}_{Y|x^*}$  has a normal distribution with mean  $(\hat{\theta}_1 x^* + \hat{\theta}_0)$  and variance shown in Eq (16.72). With  $\sigma$  unknown, substituting  $s_e$  for it, as in the previous section, leads to the result that the specific statistic,

$$t_{RL} = \frac{(\hat{\mu}_{Y|x^*} - \mu_{Y|x^*})}{s_e \sqrt{\left[ \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right]}} \quad (16.73)$$



**FIGURE 16.5:** The fitted regression line to the Density versus Ethanol Weight % data (solid line) along with the 95% confidence interval (dashed line). The confidence interval is narrowest at  $x = \bar{x}$  and widens for values further away from  $\bar{x}$ .

has a  $t$ -distribution with  $\nu = (n - 2)$  degrees of freedom. As a result, the  $(1 - \alpha) \times 100\%$  confidence interval on the regression line (mean response) at  $x = x^*$ , is:

$$\hat{\mu}_{Y|x^*} = (\hat{\theta}_1 x^* + \hat{\theta}_0) \pm t_{\alpha/2}(n - 2)s_e \sqrt{\left[ \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right]} \quad (16.74)$$

When this confidence interval is computed for all values of  $x$  of interest, the result is a confidence interval around the entire regression line. Again, as most statistical analysis software packages have the capability to compute and plot this confidence interval along with the regression line, the primary objective of this discussion is to provide the reader with a fundamental understanding of the theoretical bases for these computer outputs. For example, the 95% confidence interval for the Density-Wt% Ethanol problem in Examples 16.1 and 16.2 is shown in Fig 16.5

By virtue of the  $(x^* - \bar{x})^2$  term in Eq (16.74), a signature characteristic of these confidence intervals is that they are narrowest when  $x^* = \bar{x}$  and widen for values further away from  $\bar{x}$ .

### 16.2.5 Hypothesis Testing

For this class of problems, the hypothesis of concern is whether or not there is a real (and significant) linear functional relationship between  $x$  and  $Y$ ; i.e., whether the slope parameter,  $\theta_1 = 0$ , in which case the variation in  $Y$  is purely random around a constant mean value  $\theta_0$  (which may or may

not be zero). This translates to the following hypotheses regarding the slope parameter:

$$\begin{aligned} H_0 : \theta_1 &= 0 \\ H_a : \theta_1 &\neq 0 \end{aligned} \quad (16.75)$$

And from the preceding discussion regarding confidence intervals, the appropriate test statistic for this test, from Eq (16.64), is:

$$t_1 = \frac{\hat{\theta}_1}{s_e / \sqrt{S_{xx}}} \quad (16.76)$$

since the postulated value for the unknown  $\theta_1$  is 0; and the decision to reject or not reject  $H_0$  follows the standard two-sided  $t$ -test criteria; i.e., at the significance level  $\alpha$ ,  $H_0$  is rejected when

$$t_1 < -t_{\alpha/2}(n - 2), \text{ or } t_1 > t_{\alpha/2}(n - 2) \quad (16.77)$$

As with previous results, these conditions are identical to the  $(1 - \alpha) \times 100\%$  confidence interval on  $\theta_1$  not containing zero. When there is sufficient reason to reject  $H_0$ , the estimated regression coefficient is said to be “significant,” by which we mean that it is significantly different from zero, at the significance level  $\alpha$ .

There is nothing to prevent testing hypotheses also about the intercept parameter,  $\theta_0$ , whether or not its value is significantly different from zero. The principles are precisely as indicated above for the slope parameter; the hypotheses are,

$$\begin{aligned} H_0 : \theta_0 &= 0 \\ H_a : \theta_0 &\neq 0 \end{aligned} \quad (16.78)$$

in this case, with the test statistic (from Eq (16.65)):

$$t_0 = \frac{\hat{\theta}_0}{s_e \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \quad (16.79)$$

and the rejection criteria,

$$t_0 < -t_{\alpha/2}(n - 2), \text{ or } t_0 > t_{\alpha/2}(n - 2) \quad (16.80)$$

In addition to computing estimates of the regression coefficient and the associated standard errors, most computer programs will also compute the  $t$ -statistics and the associated  $p$ -values for each of the two coefficients.

Let us illustrate with the following example.

**TABLE 16.4:** Cranial circumference and finger lengths for 16 individuals

|                      |      |      |      |      |      |      |      |      |
|----------------------|------|------|------|------|------|------|------|------|
| Cranial Circum (cms) | 58.5 | 54.2 | 57.2 | 52.7 | 55.1 | 60.7 | 57.2 | 58.8 |
| Finger Length (cms)  | 7.6  | 7.9  | 8.4  | 7.7  | 8.6  | 8.6  | 7.9  | 8.2  |
| Cranial Circum (cms) | 56.2 | 60.7 | 53.5 | 60.7 | 56.3 | 58.1 | 56.6 | 57.7 |
| Finger Length (cms)  | 7.7  | 8.1  | 8.1  | 7.9  | 8.1  | 8.2  | 7.8  | 7.9  |

**Example 16.3: CRANIAL CIRCUMFERENCE AND FINGER LENGTH**

A once-popular exercise in the late 19<sup>th</sup> and early 20<sup>th</sup> centuries involved attempts at finding mathematical expressions that will allow one to predict, for a population of humans, some physical human attribute on the basis of a different one. The data in Table 16.4 shows the result of a classic example of such an exercise where the cranial circumference (in cms) and the length of the longest finger (in cms) of 16 individuals were determined. Postulate a linear two-parameter model as in Eq (16.19), obtain least-squares estimates of the slope and intercept, and test hypotheses that these parameters are *not* significantly different from zero. Plot the data versus the model fit and comment on the results.

**Solution:**

If  $Y$  is the cranial circumference and  $x$ , the finger length, using MINITAB to analyze this data set produces the following results:

**Regression Analysis: Cranial Circ(cm) versus Finger Length(cm)**

The regression equation is

$$\text{Cranial Circ(cm)} = 43.0 + 1.76 \text{ Finger Length(cm)}$$

| Predictor     | Coef  | SE Coef | T    | P     |
|---------------|-------|---------|------|-------|
| Constant      | 43.00 | 17.11   | 2.51 | 0.025 |
| Finger Length | 1.757 | 2.126   | 0.83 | 0.422 |

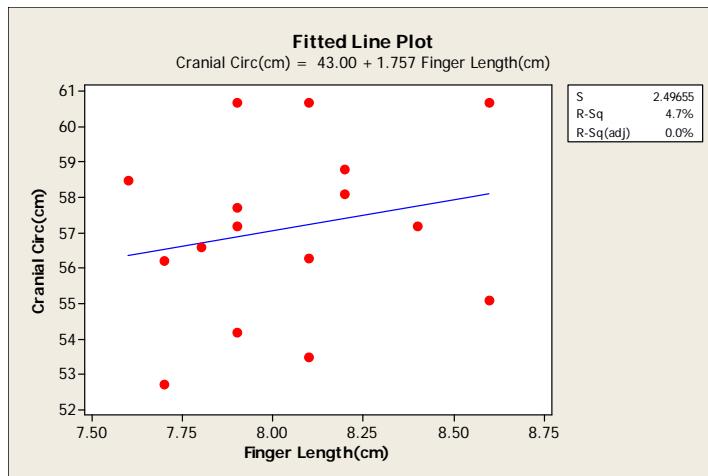
$$S = 2.49655 \quad R-\text{Sq} = 4.7\% \quad R-\text{Sq(adj)} = 0.0\%$$

Thus, the regression equation is obtained as

$$\hat{y} = 1.76x + 43.0 \quad (16.81)$$

and the model fit to the data is shown in Fig 16.6. (Again, we defer until the appropriate place, any comment on the terms included in the last line of the MINITAB output.)

It is important to note how, rather than clustering tightly around the regression line, the data shows instead a significant amount of scatter, which, at least visually, calls into question the postulated dependence of cranial circumference on finger length. This question is settled concretely by the computed  $T$  statistics for the model parameters and the



**FIGURE 16.6:** The fitted straight line to the Cranial circumference versus Finger length data. Note how the data points are widely scattered around the fitted regression line. (The additional terms included in the graph,  $S$ ,  $R\text{-}Sq$  and  $R\text{-}Sq(\text{adj})$  are discussed later)

associated  $p$ -values. The  $p$ -value of 0.025 associated with the constant (intercept parameter,  $\theta_0$ ) indicates that we must reject the null hypothesis that  $\theta_0 = 0$  in favor of the alternative that the estimated value, 43.0, is significantly different from zero, at the 5% significance level. On the other hand, the corresponding  $p$ -value associated with the  $\theta_1$ , the coefficient of  $x$ , the finger length (i.e., the regression line slope), is 0.422, indicating that there is no evidence to reject the null hypothesis. Thus, at the 5% significance level,  $\theta_1$  is *not* significantly different from zero and we therefore conclude that there is no discernible relationship between cranial circumference and finger length.

Thus, the implication of the significance of the constant term, and non-significance of the coefficient of the finger length is two-fold: (i) that cranial circumference does not depend on finger length (at least for the 16 individuals in this study), so that the observed variability is purely random, with no systematic component that can be explained by finger length; and consequently, (ii) that the cranial circumference is best characterized for this population of individuals by the mean value (43.0 cm), a value that is significantly different from zero (as one would certainly expect!).

This last example illustrates an important point about regression analysis: one can always fit any postulated model to any given set of data; the real question is: how “useful” is this model? In other words, to what extent is the implied relationship between  $x$  and  $Y$  representative of the real information contained in the data? These are very important questions that will be

answered systematically in the upcoming sections. For now, we note that at the most basic level, the hypothesis tests discussed here provide an objective assessment of the implied relationship, whether it is “real” or it is merely an artifact of random variability. Anytime we are unable to reject the null hypothesis on the slope parameter, the estimated value, and hence the model itself, are “not significant;” i.e., the real parameter value cannot be distinguished from zero.

### 16.2.6 Prediction and Prediction Intervals

A model whose parameters are confirmed as “significant” is useful for at least two things:

1. *Estimating Mean Responses:*

For a given value  $x = x^*$ , the fitted regression line provides a means of estimating the expected value of the response,  $Y|x^*$ ; i.e.,

$$E(Y|x^*) = \hat{\theta}_1 x^* + \hat{\theta}_0 \quad (16.82)$$

This is to be understood as the “least-squares” surrogate of the average of a number of replicate responses obtained when the experiment is repeated for the same fixed value  $x = x^*$ .

2. *Predicting a New Response:*

Here, the objective is slightly different: for a given  $x = x^*$ , we wish to *predict* the response observed from a *single* experiment performed at the specified value. Not surprisingly, the fitted regression line provides the best prediction,  $\hat{y}(x^*)$  as

$$\hat{y}(x^*) = \hat{\theta}_1 x^* + \hat{\theta}_0 \quad (16.83)$$

which is precisely the same as above in Eq (16.82). The difference lies not in the value themselves but in the *precision* associated with each value.

When the regression line is used as an estimator of mean (or expected) response, the precision associated with the estimate was given in the form of the variance shown in Eq (16.72), from which we developed the confidence interval around the regression line. When the regression line is used as a prediction of a yet-to-be-observed value  $Y(x^*)$ , however, the prediction error is given by:

$$E_p = Y(x^*) - \hat{Y}(x^*) \quad (16.84)$$

which, under the normality assumption, possesses the distribution  $N(0, \sigma_{E_p}^2)$ , with the variance obtained from Eq (16.84) as

$$\begin{aligned}\sigma_{E_p}^2 &= \text{Var}[Y(x^*)] + \text{Var}[\hat{Y}(x^*)] \\ &= \sigma^2 + \sigma^2 \left[ \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right] \\ &= \sigma^2 \left[ 1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right]\end{aligned}\quad (16.85)$$

where, we recall,  $\sigma^2$  as the variance of the random error component,  $\epsilon$ .

This expression differs from the expression in Eq (16.72) by the presence of the additional term, 1, making  $\sigma_{E_p}^2 > \text{Var}(\hat{\mu}_{Y|x^*})$  always. This mathematical fact is a consequence of the phenomenological fact that the prediction error is a combination of the variability inherent in determining the observed value,  $Y(x^*)$ , and the regression model error associated with  $\hat{Y}(x^*)$ , this latter quantity being the only error associated with using the regression model as an estimator of mean response.

We may now use Eq (16.85) to obtain the  $(1 - \alpha) \times 100\%$  prediction interval for  $y(x^*)$  by substituting the data estimate,  $s_e$ , for the unknown standard deviation,  $\sigma$ , and from the resulting  $t$ -distribution characteristics, i.e.,

$$y(x^*) = \hat{y}(x^*) \pm t_{\alpha/2}(n - 2)s_e \sqrt{\left[ 1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right]} \quad (16.86)$$

As with the confidence intervals around the regression line, these prediction intervals are narrowest when  $x^* = \bar{x}$  and widen for values further away from  $\bar{x}$ , but they are consistently wider than the confidence intervals.

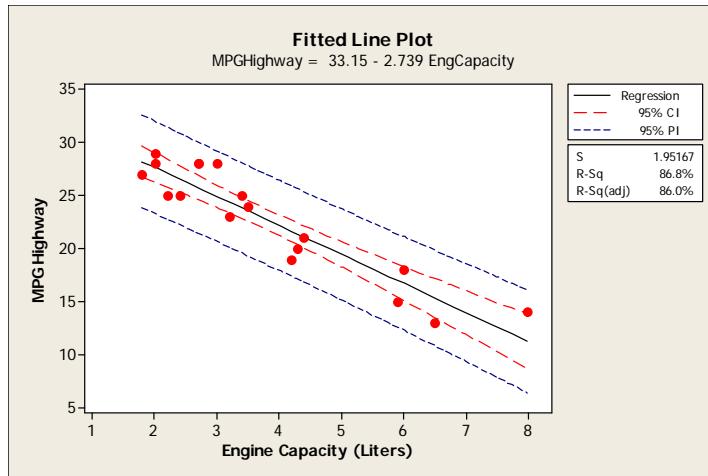
#### **Example 16.4: HIGHWAY GASOLINE MILEAGE AND ENGINE CAPACITY FOR TWO-SEATER AUTOMOBILES**

From the data shown in Table 12.5 of Chapter 12 on gasoline mileage for a collection of two-seater cars, postulate a linear two-parameter model as in Eq (16.19), for highway mileage ( $y$ ) as a function of the engine capacity,  $x$ ; obtain least-squares estimates of the parameters for all the cars, leaving out the Chevrolet Corvette and the Dodge Viper data (these cars were identified in that chapter as different from the others in the class because of the material used for their bodies). Show a plot of the fitted regression line, the 95% confidence interval and the 95% prediction interval.

#### **Solution:**

Using MINITAB for this problem produces the following results:

**Regression Analysis: MPGHighway versus EngCapacity**  
The regression equation is



**FIGURE 16.7:** The fitted straight line to the Highway MPG versus Engine Capacity data of Table 12.5 (leaving out the two “inconsistent” data points) along with the 95% confidence interval (long dashed line) and the 95% prediction interval (short dashed line). (Again, the additional terms included in the graph,  $S$ ,  $R\text{-Sq}$  and  $R\text{-Sq}(\text{adj})$  are discussed later).

| <u>MPGHighway = 33.2 - 2.74 EngCapacity</u> |         |         |        |       |  |
|---|---------|---------|--------|-------|--|
| Predictor                                   | Coef    | SE Coef | T      | P     |  |
| Constant                                    | 33.155  | 1.110   | 29.88  | 0.000 |  |
| EngCapacity                                 | -2.7387 | 0.2665  | -10.28 | 0.000 |  |

$$S = 1.95167 \quad R\text{-Sq} = 86.8\% \quad R\text{-Sq}(\text{adj}) = 86.0\%$$

Thus, with Highway MPG as  $y$ , and Engine Capacity as  $x$ , the fitted regression line equation is

$$\hat{y} = -2.74x + 33.2 \quad (16.87)$$

and, since the  $p$ -values associated with each parameter are both zero to three decimal places, we conclude that these parameters are “significant.” The implication is that for every liter increase in engine capacity, the average two-seater car is expected to lose about 2 and  $3/4$  miles per gallon on the highway. (As before, we defer until later any comment on the terms in the last line of the MINTAB output.)

The model fit to the data is shown in Fig 16.7 along with the required 95% confidence interval (CI) and the 95% prediction interval (PI). Note how much wider the PI is than the CI at every value of  $x$ .

### 16.2.7 Coefficient of Determination and the F-Test

Beyond hypotheses tests to determine the significance of *individual* estimated parameters, other techniques exist for assessing the *overall* effectiveness of the regression model, based on measures of how much of the total variability in the data has been captured (or explained) by the model.

#### Orthogonal Decomposition of Variability

The total variability present in the data, represented by  $\sum_{i=1}^n (y_i - \bar{y})^2$ , and defined as  $S_{yy}$  in Eq (16.32), may be rearranged as follows, merely by adding and subtracting  $\hat{y}_i$ :

$$\begin{aligned} S_{yy} &= \sum_{i=1}^n (y_i - \bar{y})^2 \\ &= \sum_{i=1}^n [(y_i - \hat{y}_i) - (\bar{y} - \hat{y}_i)]^2 \end{aligned} \quad (16.88)$$

Upon expanding and simplifying (see Exercise 16.9), one obtains the very important expression:

$$\begin{aligned} \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ \text{or } S_{yy} &= SS_R + SS_E \end{aligned} \quad (16.89)$$

where we have recalled that the second term on the RHS of the equation is the residual error sum of squares defined in Eq (16.59), and have introduced the term  $SS_R$  to represent the regression sum of squares,

$$SS_R = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad (16.90)$$

a measure of the variability represented in the regression line's estimate of the mean response. The expression in Eq (16.89) represents a decomposition of the total variability in the data into two components: the variability captured by the regression model,  $SS_R$ , and what is left in the residual error,  $SS_E$ . In fact, we had actually encountered this expression earlier, in Eq (16.61), where what we now refer to as  $SS_R$  had earlier been presented as  $\hat{\theta}_1 S_{xy}$ . If the data vector is represented as  $\mathbf{y}$ , the corresponding vector of regression model estimates as  $\hat{\mathbf{y}}$ , and the vector of residual errors between these two as  $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ , then observe that

$$(\mathbf{y} - \bar{\mathbf{y}}) = (\hat{\mathbf{y}} - \bar{\mathbf{y}}) + \mathbf{e} \quad (16.91)$$

But from the definition of vector Euclidian norms,

$$\|(\mathbf{y} - \bar{\mathbf{y}})\|^2 = S_{yy} \quad (16.92)$$

$$\|(\hat{\mathbf{y}} - \bar{\mathbf{y}})\|^2 = SS_R \quad (16.93)$$

$$\|\mathbf{e}\|^2 = SS_E \quad (16.94)$$

with the very important implication that, as a result of the vector representation in Eq (19.11), the expression in Eq (16.89) is an orthogonal decomposition of the data variance vector reminiscent of Pythagoras' Theorem. (If the vector sum in Eq (19.11) holds simultaneously as the corresponding sums of squares expression in Eq (16.89), then the vector  $(\hat{\mathbf{y}} - \bar{\mathbf{y}})$  must be orthogonal to the vector  $\mathbf{e}$ .)

Eq (16.89) is in fact known as the *analysis of variance (ANOVA) identity*; and it plays a central role in statistical inference that transcends the restricted role observed here in regression analysis. We shall have cause to revisit this subject in our discussion of the design of experiments in upcoming chapters. For now, we use it to assess the effectiveness of the overall regression model (as a single entity purporting to represent the information contained in the data), first in the form of the coefficient of determination, and later as the basis for an *F*-test of significance. This latter exercise will constitute a preview of an upcoming, more general discussion of ANOVA.

### $R^2$ , The Coefficient of Determination

Let us now consider the ratio defined as:

$$R^2 = \frac{SS_R}{S_{yy}} \quad (16.95)$$

which represents the proportion of the total data variability (around the mean  $\bar{y}$ ) that has been captured by the regression model; its complement,

$$1 - R^2 = SS_E/S_{yy} \quad (16.96)$$

is the portion left unexplained by the regression model. Observe that  $0 \leq R^2 \leq 1$ , and that if a model adequately captures the relevant information contained in a data set, what will be left unexplained as random variation should be comparatively small, so that the  $R^2$  value will be close to 1. Conversely, a value close to zero indicates a model that is inadequate in capturing the important variability present in the data.  $R^2$  is therefore known as the coefficient of determination; it is a direct measure of the quality of fit provided by the regression model.

Although not directly relevant yet at this point (where we are still discussing the classical two-parameter model), it is possible to improve a model fit by introducing additional parameters. Under such circumstances, the improvement in  $R^2$  may come at the expense of over-fitting (as discussed more

fully later). A somewhat more judicious assessment of model adequacy requires adjusting the value of  $R^2$  to reflect the number of parameters that have been used by the model to capture the variability.

By recasting the expression in Eq (16.95) in the equivalent form:

$$R^2 = 1 - \frac{SS_E}{S_{yy}} \quad (16.97)$$

rather than base the metric on the indicated absolute sums of squares, consider using the mean sums of squares instead. In other words, instead of the total residual error sum of squares,  $SS_E$ , we employ instead the mean residual error sum of squares,  $SS_E/(n-p)$  where  $p$  is the number of parameters in the model and  $n$  is the total number of experimental data points; also instead of the total data sum of squares,  $S_{yy}$ , we employ instead the data variance  $S_{yy}/(n-1)$ . The resulting quantity, known as  $R_{adj}^2$ , and defined as:

$$R_{adj}^2 = 1 - \frac{SS_E/(n-p)}{S_{yy}/(n-1)} \quad (16.98)$$

is similar to the coefficient of determination,  $R^2$ , but it is adjusted for the the number of parameters contained in the model. It penalizes models that achieve decent values of  $R^2$  via the use of an excessive number of parameters. Relatively high values of  $R^2$  and  $R_{adj}^2$  that are also comparable in magnitude indicate a model that is quite adequate: the variability in the data has been captured adequately without using an excessive number of parameters.

All software packages that carry out regression analysis routinely compute  $R^2$  and  $R_{adj}^2$ , sometimes presented not as fractions (as indicated above), but multiplied by 100%. In fact, all the examples and fitted regression line plots encountered thus far in this chapter have shown these values (in percentage form) but we had to defer commenting on them until now. We are only now in a position for such a discussion.

In Figs 16.4, 16.5, 16.6 and 16.7, the value shown for  $S$  is the square root of the mean residual sum of squares, i.e.,  $\sqrt{SS_E/(n-2)}$ , an estimate of the unknown data standard deviation,  $\sigma$ ; this is accompanied by values for  $R^2$  and  $R_{adj}^2$ . Thus, in the Density-Ethanol weight percent regression model, (Fig 16.4), both  $R^2$  and  $R_{adj}^2$  are reported as 99.8%, indicating a model that appears to have explained virtually all the variability in the data, with very little left by way of the residual error (as indicated by the very small value of  $S$ ). The exact opposite is the case with the Cranial circumference versus Finger length regression model:  $R^2$  is an incredibly low 4.7% and the  $R_{adj}^2$  vanishes entirely (a “perfect” 0.00%), indicating that (a) the model has explained very little of the variability in the data, and (b) when penalized for the parameters employed in achieving even the less than 5% variability captured, the inadequacy of the model is seen to be total. The residual data standard deviation,  $S$ , is almost 2.5. With the Highway gas mileage versus Engine capacity regression model, the  $R^2$  value is reasonably high at 86.8%, with an adjusted value of 86% that

is essentially unchanged; the residual standard of  $S = 1.95$  is also reasonable. The indication is that while there is still some unexplained variability left, the regression model captures a significant amount of the variability in the data and provides a reasonable mathematical explanation of the information contained in the data.

#### F-test for Significance of Regression

Let us return to the ANOVA expression in Eq (16.89);

$$\begin{aligned}\sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ \text{i.e } S_{yy} &= SS_R + SS_E\end{aligned}$$

and note the following:

1. The total sum of squares,  $S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$ , has  $(n - 1)$  degrees of freedom; the error sum of squares,  $SS_E = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ , has  $(n - 2)$  degrees of freedom, and the regression sum of squares,  $SS_R = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ , has 1 degree of freedom.
2. One informal way to confirm this fact is as follows: (i) of the  $n$  independent units of information in the raw data,  $y_i; i = 1, 2, \dots, n$ , one “degree of freedom” is “tied up” in obtaining the average,  $\bar{y}$ , so that  $(y_i - \bar{y})$  will have  $(n - 1)$  degrees of freedom left (i.e., there are now only  $(n - 1)$  independent quantities in  $(y_i - \bar{y})$ ); (ii) similarly, 2 “degrees of freedom” are “tied up” in obtaining the response estimate  $\hat{y}$  (via the two parameters,  $\hat{\theta}_0$  and  $\hat{\theta}_1$ ), so that  $(y_i - \hat{y}_i)$  has  $(n - 2)$  degrees of freedom left; and finally (iii) while  $\hat{y}_i$  “ties up” 2 degrees of freedom,  $\bar{y}$  “ties up” one, so that  $(\hat{y}_i - \bar{y})$  has one degree of freedom left.
3. The implication is therefore that, in addition to representing a decomposition of variability, since it is also true that

$$(n - 1) = 1 + (n - 2) \quad (16.99)$$

Eq (16.89) also represents a concurrent decomposition of the degrees of freedom associated with each sum of squares.

Finally, from the following results (given without proof; e.g., Eq (16.62)),

$$\begin{aligned}E(SS_E) &= (n - 2)\sigma^2 \\ E(SS_R) &= \hat{\theta}_1^2 S_{xx} + \sigma^2\end{aligned} \quad (16.100)$$

we arrive at the following conclusions: Under the null hypothesis  $H_0 : \theta_1 = 0$ , these two equations suggest  $SS_E/(n - 2)$  and  $SS_R/1$  (respectively the error mean square,  $MS_E$  and the regression mean square,  $MS_R$ ) as two separate

**TABLE 16.5:** ANOVA Table for Testing Significance of Regression

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Square | F                   | p |
|---------------------|----------------|--------------------|-------------|---------------------|---|
| Regression          | $SS_R$         | 1                  | $MS_R$      | $\frac{MS_R}{MS_E}$ |   |
| Error               | $SS_E$         | $(n - 2)$          | $MS_E$      |                     |   |
| Total               | $S_{yy}$       | $(n - 1)$          |             |                     |   |

and distinct estimators of  $\sigma^2$ . Furthermore, under the normality assumption for  $y_i$ , then the statistic

$$F = \frac{SS_R/1}{SS_E/(n-2)} \quad (16.101)$$

will possess an  $F(\nu_1, \nu_2)$  distribution, with  $\nu_1 = 1$ , and  $\nu_2 = (n - 2)$ , if  $H_0$  is true that  $\theta_1 = 0$ . However, if  $H_0$  is not true, then the numerator in Eq (16.101) will be inflated by the term  $\hat{\theta}_1^2 S_{xx}$  as indicated in Eq (16.100). Hence, at the significance level of  $\alpha$ , we reject  $H_0$  (that the regression as a whole is *not* significant), when the actual computed statistic

$$f > f_\alpha(\nu_1, \nu_2) \quad (16.102)$$

where  $f_\alpha(\nu_1, \nu_2)$  is the usual  $F(\nu_1, \nu_2)$ -distribution variate with upper tail area  $\alpha$ . Equivalently, one computes the *p*-value associated with the computed  $f$ , as

$$p = P(F \geq f) \quad (16.103)$$

and reject or fail to reject the null hypothesis on the basis of the actual *p*-value.

These results are typically presented in what is referred to as an ANOVA Table as shown in Table 16.5. They are used to carry out *F*-tests for the significance of the entire regression model as a single entity; if the resulting *p*-value is low, we reject the null hypothesis and conclude that the regression is significant, i.e., the relationship implied by the regression model is meaningful. Alternatively, if the *p*-value exceeds a pre-specified threshold, (say, 0.05), we fail to reject the null hypothesis and conclude that the regression model is not significant — that the implied relationship is purely random.

All computer programs that perform regression analysis produce such ANOVA tables. For example, the MINITAB output for Example 16.4 above (involving the regression model relating engine capacity to the highway mpg rating) includes the following ANOVA table.

#### Analysis of Variance

| Source         | DF | SS     | MS     | F      | P     |
|----------------|----|--------|--------|--------|-------|
| Regression     | 1  | 402.17 | 402.17 | 105.58 | 0.000 |
| Residual Error | 16 | 60.94  | 3.81   |        |       |
| Total          | 17 | 463.11 |        |        |       |

The indicated  $p$ -value of 0.000 implies that we must reject the null hypothesis and conclude that the regression model is “significant.”

On the other hand, the ANOVA table produced by MINITAB for the cranial circumference versus finger length regression problem of Example 16.3 is as shown below:

#### Analysis of Variance

| Source         | DF | SS     | MS    | F    | P     |
|----------------|----|--------|-------|------|-------|
| Regression     | 1  | 4.259  | 4.259 | 0.68 | 0.422 |
| Residual Error | 14 | 87.259 | 6.233 |      |       |
| Total          | 15 | 91.517 |       |      |       |

In this case, the  $p$ -value associated with the  $F$ -test is so high (0.422) that we reject the null hypothesis and conclude that the regression is *not* significant.

Of course, these conclusions agree perfectly with our earlier conclusions concerning each of these problems.

In general, we tend to de-emphasize these ANOVA-based  $F$ -tests for significance of the regression. This is for the simple reason that they are coarse tests of the *overall* regression model, adding little or nothing to the individual  $t$ -tests presented earlier for each parameter. These individual parameter tests are preferred because they are finer-grained.

From this point on, we will no longer refer to these ANOVA tests of significance for regression. Nevertheless, these same concepts take center stage in Chapter 19 where they are central to analysis of designed experiments.

#### 16.2.8 Relation to the Correlation Coefficient

In Chapter 5, we defined the correlation coefficient between two jointly distributed random variables  $X$  and  $Y$  as:

$$\rho = \frac{E[XY]}{\sigma_X \sigma_Y} \quad (16.104)$$

The sample version, obtained from data, known as the Pearson product moment correlation coefficient, is:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (16.105)$$

If we now recall the expressions in Eqs (16.31–16.33), we immediately obtain, in the context of regression analysis:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}} \sqrt{S_{yy}}} \quad (16.106)$$

And now, in terms of the slope parameter estimate, we obtain from Eq (16.37), first that

$$r = \hat{\theta}_1 \frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}} \quad (16.107)$$

an expression we shall return to shortly. For now, let us return to the expression for  $R^2$  in Eq (16.97); if we introduce Eq (16.61) for  $SS_E$ , we obtain

$$R^2 = 1 - \frac{S_{yy} - \hat{\theta}_1 S_{xy}}{S_{yy}} = \hat{\theta}_1 \frac{S_{xy}}{S_{yy}} \quad (16.108)$$

Upon introducing Eq (16.37) for  $\hat{\theta}_1$ , we obtain the result that:

$$R^2 = \frac{(S_{xy})^2}{S_{xx} S_{yy}} \quad (16.109)$$

which, when compared with Eq (16.106) establishes the important result that  $R^2$ , the coefficient of determination, is the square of the sample correlation coefficient,  $r$ ; i.e.,

$$R^2 = r^2. \quad (16.110)$$

### 16.2.9 Mean-Centered Model

As obtained previously, the estimated observation from the regression model is given by

$$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x \quad (16.111)$$

and, from a rearrangement of Eq (16.38) used to estimate  $\hat{\theta}_0$ , we obtain

$$\bar{y} = \hat{\theta}_0 + \hat{\theta}_1 \bar{x} \quad (16.112)$$

If we now subtract the latter equation from the former, we obtain:

$$(\hat{y} - \bar{y}) = \hat{\theta}_1 (x - \bar{x}) \quad (16.113)$$

which is a “mean-centered” version of the regression model.

If we now rearrange Eq (16.107) to express  $\hat{\theta}_1$  in terms of  $r$  and introduce it into Eq (16.113), we obtain:

$$(\hat{y} - \bar{y}) = r \left( \frac{s_y}{s_x} \right) (x - \bar{x}) \quad (16.114)$$

where  $s_y = \sqrt{S_{yy}}/(n-1)$  and  $s_x = \sqrt{S_{xx}}/(n-1)$ , are, respectively sample estimates of the data standard deviation for  $y$  and for  $x$ . Alternatively, Eq (16.114) could equivalently be written as

$$(\hat{y} - \bar{y}) = \sqrt{R^2 \left( \frac{S_{yy}}{S_{xx}} \right)} (x - \bar{x}) \quad (16.115)$$

This equation provides the clearest indication of the impact of  $R^2$  on how “strongly” the mean-centered value of the predictor,  $x$ , is connected by the

model to — and hence can be used to estimate — the mean-centered response. Observe that in the density and weight percent ethanol example, with  $R^2 = 0.998$ , the connection between the predictor and response estimate is particularly strong; with the cranial circumference-finger length example, the connection is extremely weak, and the best estimate of the response (cranial circumference) for any value of the predictor (finger length), is essentially the mean value,  $\bar{y}$ .

### 16.2.10 Residual Analysis

While the statistical significance of the estimated parameters gives us some information about the usefulness of a model, and while the  $R^2$  and  $R_{adj}^2$  values provide a measure of how much of the data's variability has been captured by the overall model, how well the model represents the data is most directly determined from the difference between the actual observation,  $y_i$ , and the model estimate,  $\hat{y}_i$ , i.e.,  $e_i = y_i - \hat{y}_i, i = 1, 2, \dots, n$ . These quantities, identified earlier as residual errors (or simply as “residuals”), provide  $n$  different samples of how closely the model matches the data. If the model's representation of the data is adequate, the residuals should be nothing but purely random variation. Any departure from pure random variation in the residuals is an indication of some form or the other of model inadequacy. Residual analysis therefore allows us to do the following:

1. Check model assumptions, specifically that  $\epsilon \sim N(0, \sigma^2)$ , with  $\sigma^2$  constant;
2. Identify any “left over” systematic structural discrepancies between the model and data; and
3. Identify which data points might be inconsistent with the others in the set.

By formulation, the least squares estimation technique *always* produces a model for which  $\sum_{i=1}^n e_i = 0$ , making the desired zero mean characteristics of the model error sequence a non-issue. Residual analysis is therefore concerned mostly with the following activities:

1. Formal and informal tests of normality of  $e_i$ ;
2. Graphical/visual evaluation of residual patterns;
3. Graphical/visual and numerical evaluation of individual residual magnitude.

Other than formal normality tests (which involve techniques discussed in the next chapter) residual analysis is more of an art involving various graphical plots. When there is sufficient data, histograms of the residuals provide great visual clues regarding normality quite apart from what formal tests show. In

**TABLE 16.6:** Thermal conductivity measurements at various temperatures for a metal

| $k$ (W/m-°C) | Temperature (°C) |
|--------------|------------------|
| 93.228       | 100              |
| 92.563       | 150              |
| 99.409       | 200              |
| 101.590      | 250              |
| 111.535      | 300              |
| 115.874      | 350              |
| 119.390      | 400              |
| 126.615      | 450              |

many cases, however, available data is usually modest. Regardless of the data size, plots of the residuals themselves versus the fitted value,  $\hat{y}$ , or versus data order, or versus  $x$ , are not only capable of indicating model adequacy; they also provide clues about the nature of the implied model inadequacy.

It is often recommended that residual plots be based not on the residual themselves, but on the standardized residual,

$$e_i^* = \frac{e_i}{s_e} \quad (16.116)$$

where,  $s_e$ , as we recall, is the estimate of  $\sigma$ , the data standard deviation. This is because if the residuals are truly normally distributed, then  $-2 < e_i^* < 2$  for approximately 95% of the standardized residuals; and some 99% should lie between  $-3$  and  $3$ . As a general rule-of-thumb, therefore,  $|e_i^*| > 3$  indicates a value that is considered a potential “outlier” because it is inconsistent with the others.

When a model appears inadequate, the recommendation is to look for clues within the residuals for what to do next. The following examples illustrate residual analysis for practical problems in engineering.

**Example 16.5: TEMPERATURE DEPENDENCE OF THERMAL CONDUCTIVITY**

To characterize how the thermal conductivity,  $k$  (W/m-°C), of a metal varies with temperature, eight independent experiments were performed at the temperatures,  $T^\circ C$ , shown in Table 16.6 along with the measured thermal conductivities. A two-parameter model as in Eq 16.19 has been postulated for the relationship between  $k$  and  $T$ . Obtain a least-squares estimate of the parameters and evaluate the model fit to the data.

**Solution:**

We use MINITAB and obtain the following results:

**Regression Analysis: k versus Temperature**  
The regression equation is

| $k = 79.6 + 0.102 \text{ Temperature}$ |          |          |       |       |
|--|----------|----------|-------|-------|
| Predictor                              | Coef     | SE Coef  | T     | P     |
| Constant                               | 79.555   | 2.192    | 36.29 | 0.000 |
| Temperature                            | 0.101710 | 0.007359 | 13.82 | 0.000 |

$$S = 2.38470 \quad R-\text{Sq} = 97.0\% \quad R-\text{Sq}(\text{adj}) = 96.4\%$$

Therefore, as before, representing the thermal conductivity as  $y$ , and Temperature as  $x$ , the fitted regression line equation is

$$\hat{y} = 0.102x + 79.6 \quad (16.117)$$

The  $p$ -values associated with each parameter is zero, implying that both parameters are significantly different from zero. The estimate of the data standard deviation is shown as  $S$ ; and the  $R^2$  and  $R_{\text{adj}}^2$  values indicate that the model captures a reasonable amount of the variability in the data.

The actual model fit to the data is shown in the top panel of Fig 16.8 while the standardized residuals versus fitted value,  $\hat{y}_i$ , is shown in the bottom panel. With only 8 data points, there are not enough residuals for a histogram plot. Nevertheless, upon visual examination of the residual plots, there appears to be no discernible pattern, nor is there any reason to believe that the residuals are anything but purely random. Note that no standardized residual value exceeds  $\pm 2$ .

The model is therefore considered to provide a reasonable representation of how the thermal conductivity of this metal varies with temperature.

The next example illustrates a practical circumstance where the residuals not only expose the inadequacy of a linear regression model, but also provide clues concerning how to rectify the inadequacy.

#### Example 16.6: BOILING POINT OF HYDROCARBONS

It has been proposed to represent with a linear two-parameter model, the relationship between the number of carbon atoms in the hydrocarbon compounds listed in Table 16.1 and the respective boiling points. Evaluate a least-squares fit of this model to the data.

#### Solution:

Using MINITAB produces the following results for this problem:

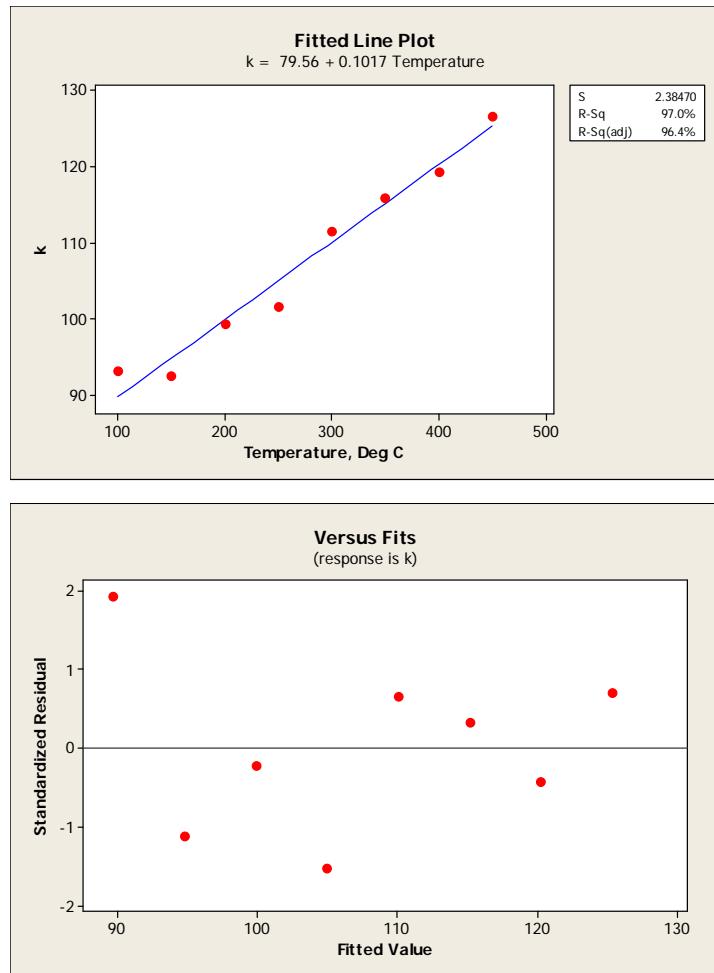
#### Regression Analysis: Boiling Point versus n

The regression equation is

$$\text{Boiling Point} = -172.8 + 39.45 n$$

| Predictor | Coef    | SE Coef | T      | P     |
|-----------|---------|---------|--------|-------|
| Constant  | -172.79 | 13.26   | -13.03 | 0.000 |
| n         | 39.452  | 2.625   | 15.03  | 0.000 |

$$S = 17.0142 \quad R-\text{Sq} = 97.4 \% \quad R-\text{Sq}(\text{adj}) = 97.0\%$$



**FIGURE 16.8:** Modeling the temperature dependence of thermal conductivity: Top: Fitted straight line to the Thermal conductivity ( $k$ ) versus Temperature ( $T^{\circ}\text{C}$ ) data in Table 16.6; Bottom: standardized residuals versus fitted value,  $\hat{y}_i$ .

Therefore, as before, the fitted regression line equation is

$$\hat{y} = 39.45x - 172.8 \quad (16.118)$$

with the hydrocarbon compound boiling point as  $y$ , and the number of carbon atoms it contains as  $x$ .

We notice, once again, that for this model, the parameter values are all significantly different from zero because the  $p$ -values are zero in each case; furthermore, the  $R^2$  and  $R_{adj}^2$  values are quite good. The error standard deviation is obtained as  $S = 17.0142$ . By themselves, nothing in these results seem out of place; one might even be tempted by the excellent  $R^2$  and  $R_{adj}^2$  values to declare that this is a very good model. However, the model fit to the data, shown in the top panel of Fig 16.9, tells a different story; and the normalized residuals versus fitted value,  $\hat{y}_i$ , shown in the bottom panel, is particularly revealing. The model fit shows a straight line model that very consistently overestimates BP at the extremes and underestimates it in the middle. The standardized residual versus model fit quantifies this under- and over-estimation and shows a clear “left over” quadratic structure.

The implication clearly is that while approximately 97% of the relationship between  $n$ , the number of carbon atoms, and the hydrocarbon BP has been captured by a linear relationship, there remains an unexplained, possibly quadratic, component that is clearly discernible. The suggestion: consider a revised model of the type

$$Y = \theta_0 + \theta_1 x + \theta_2 x^2 \quad (16.119)$$

Such a model is a bit more complicated, but the residual structure seems to suggest that the additional term is warranted. How to obtain a model of this kind is discussed shortly.

We revisit the problem illustrated in this example after completing a discussion of more complicated regression models in the upcoming sections.

### 16.3 “Intrinsically” Linear Regression

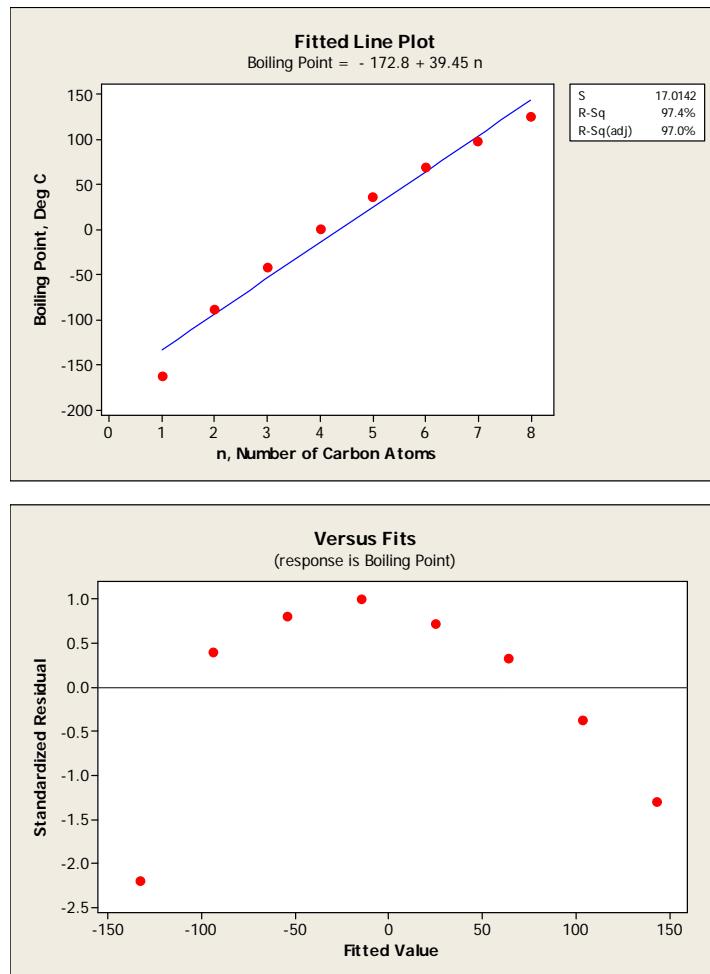
#### 16.3.1 Linearity in Regression Models

In estimating the vector of parameters  $\boldsymbol{\theta}$  contained in the general regression model,

$$Y = g(x; \boldsymbol{\theta}) + \epsilon$$

it is important to clarify what the term “linear” in linear regression refers to. While it is true that the model:

$$Y = \theta_0 + \theta_1 x + \epsilon$$



**FIGURE 16.9:** Modeling the dependence of the boiling points (BP) of hydrocarbon compounds in Table 16.1 on the number of carbon atoms in the compound: Top: Fitted straight line of BP versus  $n$ , the number of carbon atoms; Bottom: standardized residuals versus fitted value  $\hat{y}_i$ . Notice the distinctive quadratic structure "left over" in the residuals exposing the linear model's over-estimation at the extremes and the under-estimation in the middle.

is a linear equation because it represents a straight line relationship between  $Y$  and  $x$ , what is actually of relevance in regression analysis is that this functional form is linear with respect to the unknown parameter vector  $\boldsymbol{\theta} = (\theta_0, \theta_1)$ . It must be kept in mind in regression, that  $x$  is known and given; the parameters  $\boldsymbol{\theta}$  are the unknowns to be determined by the regression exercise.

Thus, what is of importance in determining whether a regression problem is linear or not is the functional form of  $g(x; \boldsymbol{\theta})$  with respect to the vector of parameters  $\boldsymbol{\theta}$ , *not with respect to x*. For example, if the regression model is given as

$$Y = \theta_1 x^n + \epsilon, \quad (16.120)$$

clearly this is a nonlinear function of  $x$ ; however, so long as  $n$  is known, for any given value of  $x$ ,  $x^n$  is also known, say  $x^n = z$ ; this equation is therefore exactly equivalent to

$$Y = \theta_1 z + \epsilon, \quad (16.121)$$

which is clearly linear. Thus, even though nonlinear in  $x$ , Eq (16.120) nevertheless represents a linear regression problem because the model equation is linear in the unknown parameter  $\theta$ . On the other hand, the model representing how the concentration  $C(t)$  of a reactant undergoing a first order kinetic reaction in a batch reactor changes with time,

$$C(t) = \theta_0 e^{-\theta_1 t} \quad (16.122)$$

with time,  $t$ , as the independent variable, along with  $\theta_0$  and  $\theta_1$  respectively as the unknown initial concentration and kinetic reaction rate constant, represents a truly *nonlinear* regression model. This is because one of the unknown parameters,  $\theta_1$ , enters the model nonlinearly; the model is linear in the other parameter  $\theta_0$ .

As far as regression is concerned, therefore, whether the problem at hand is linear or nonlinear, depends on whether the parameter sensitivity function,

$$S_{\theta_i} = \frac{\partial g}{\partial \theta_i} \quad (16.123)$$

is a function of  $\theta_i$  or not. For linear regression problems  $S_{\theta_i}$  is independent of  $\theta_i$  for all  $i$ ; the defining characteristics of nonlinear regression is that  $S_{\theta_i}$  depends on  $\theta_i$  for at least one  $i$ .

#### **Example 16.7: LINEAR VERSUS NONLINEAR REGRESSION PROBLEMS**

Which of the following three models presents a linear or nonlinear regression problem in estimating the indicated unknown parameters,  $\theta_i$ ?

$$(1) Y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \epsilon \quad (16.124)$$

$$(2) Y = \theta_1 \theta_2^x + \epsilon \quad (16.125)$$

$$(3) Y = \theta_1 e^{x_1} + \theta_2 \ln x_2 + \frac{\theta_3}{x_3} + \epsilon \quad (16.126)$$

**Solution:**

Model (1) presents a linear regression problem because each of the sensitivities,  $\mathcal{S}_{\theta_0} = 1$ ;  $\mathcal{S}_{\theta_1} = x$ ;  $\mathcal{S}_{\theta_2} = x^2$ ; and  $\mathcal{S}_{\theta_3} = x^3$ , is free of the unknown parameter on which it is based. i.e.,  $\mathcal{S}_{\theta_i}$  is not a function of  $\theta_i$  for  $i = 0, 1, 2, 3$ . In fact, all the sensitivities are entirely free of all parameters.

Model (2) on the other hand presents a nonlinear regression problem:  $\mathcal{S}_{\theta_1} = \theta_2^x$  does not depend on  $\theta_1$ , but

$$\mathcal{S}_{\theta_2} = \theta_1 x \theta_2^{x-1} \quad (16.127)$$

depends on  $\theta_2$ . Thus, while this model is linear in  $\theta_1$  (because the sensitivity to  $\theta_1$  does not depend on  $\theta_1$ ), it is nonlinear in  $\theta_2$ ; therefore, it presents a nonlinear regression problem.

Model (3) presents a linear regression problem:  $\mathcal{S}_{\theta_1} = e^{x_1}$ ;  $\mathcal{S}_{\theta_2} = \ln x_2$ , are both entirely free of unknown parameters.

### 16.3.2 Variable Transformations

A restricted class of truly nonlinear regression models may be converted to linear models via appropriate variable transformations; linear regression analysis can then be carried out in terms of the transformed variables. For example, observe that even though the reactant concentration model in Eq (16.122) has been identified as nonlinear in the parameters, a logarithmic transformation results in:

$$\ln C(t) = \ln \theta_0 - \theta_1 t \quad (16.128)$$

In this case, observe that if we now let  $Y = \ln C(t)$ , and let  $\theta_0^* = \ln \theta_0$ , then Eq (16.128) represents a linear regression model.

Such cases abound in chemical engineering. For example, the equilibrium vapor mole fraction,  $y$ , as a function of liquid mole fraction,  $x$ , of a compound with relative volatility  $\alpha$  is given by the expression:

$$y = \frac{\alpha x}{1 + (\alpha - 1)x} \quad (16.129)$$

It is an easy exercise to show that by inverting this equation, we obtain:

$$\frac{1}{y} = \theta \frac{1}{x} + (1 - \theta) \quad (16.130)$$

so that  $1/y$  versus  $1/x$  produces a linear regression problem.

Such models are said to be “intrinsically” linear because while non linear in their original variables, they are linear in a different set of transformed variables; the task is to find the required transformation. Nevertheless, the careful reader would have noticed something missing from these model equations: we have carefully avoided introducing the error terms,  $\epsilon$ . This is for the simple

reason that in virtually all cases, if the error term is additive, then even the obvious transformations are no longer possible. For example, if each actual concentration measurement,  $C(t_i)$ , observed at time  $t_i$  has associated with it the additive error term  $\epsilon_i$ , then Eq (16.122) must be rewritten as

$$C(t_i) = \theta_0 e^{-\theta_1 t_i} + \epsilon_i \quad (16.131)$$

and the logarithmic transformation is no longer possible.

Under such circumstances, most “practitioners” will suspend the addition of the error term until after the function has been appropriately transformed; i.e., instead of writing the model as in Eq (16.131), it would be written as:

$$\ln C(t_i) = \ln \theta_0 - \theta_1 t_i + \epsilon_i \quad (16.132)$$

But this implies that the error is multiplicative in the original variable. It is important, before taking such a step, to take time to consider whether such a multiplicative error structure is reasonable or not.

Thus, in employing transformations to deal with these so-called intrinsically linear models, the most important issue lies in determining the proper error structure. Such transformations should be used with care; alternatively, the parameter estimates obtained from such an exercise should be considered as approximations that may require further refinement by using more advanced nonlinear regression techniques. Notwithstanding, many engineering problems involving models of this kind have benefited from the sort of linearizing transformations discussed here.

## 16.4 Multiple Linear Regression

In many cases, the response variable  $Y$  depends on several independent variables,  $x_1, x_2, \dots, x_m$ . Under these circumstances, the simplest possible regression model is:

$$Y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_m x_m + \epsilon \quad (16.133)$$

with  $m$  independent predictor variables,  $x_i$ ;  $i = 1, 2, \dots, m$ , and  $m+1$  unknown parameters,  $\boldsymbol{\theta} = (\theta_0, \theta_1, \dots, \theta_m)$ , the “regression coefficients.” Eq (16.133) represents a multiple linear regression model. An example is when  $Y$ , the conversion obtained from a catalytic process depends on the temperature,  $x_1$ , and pressure,  $x_2$ , at which the reactor is operated, according to:

$$Y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \epsilon \quad (16.134)$$

Just as the expected value of the response in Eq (16.19) represents a straight line, the expected value in Eq (16.133) represents an  $m$ -dimensional hyperplane.

However, there is no reason to restrict the model to the form in Eq (16.133). With more than one independent variable, it is possible for the response variable to depend on higher order powers of, as well as interactions between, some of the variables: for example, a better representation of the relationship between yield and temperature and pressure might be:

$$Y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_{11} x_1^2 + \theta_{22} x_2^2 + \theta_{12} x_1 x_2 + \epsilon \quad (16.135)$$

Such models as these are often specifically referred to as “response surface” models and we shall have cause to revisit them later on in upcoming chapters. For now, we note that in general, most multiple regression models are, for the most part, justified as approximations to the more general expression,

$$Y = g(x_1, x_2, \dots, x_m; \boldsymbol{\theta}) + \epsilon \quad (16.136)$$

where neither the true form of  $g(x_1, x_2, \dots, x_m; \boldsymbol{\theta})$ , nor the parameters,  $\boldsymbol{\theta}$  are known. If  $g(\cdot)$  is analytic, then by taking a Taylor series expansion and truncating after a pre-specified number of terms, the result will be a multiple regression model. The multiple regression function is therefore often justified as a Taylor series approximation of an unknown, and perhaps more complex, function. For example, Eq (16.135) arises when the Taylor expansion only goes up to the second order.

Regardless of what form the regression model takes, keep in mind that so long as the values of the independent variables are known, all such models can always be recast in the form shown in Eq (16.133). For example, even though there are two actual predictor variables  $x_1$  and  $x_2$  in the model in Eq (16.135), if we define “new” variables  $x_3 = x_1^2; x_4 = x_2^2; x_5 = x_1 x_2$ , then Eq (16.135) immediately becomes like Eq (16.133) with  $m = 5$ . Thus, it is without loss of generality that we consider Eq (16.133) as the general multiple regression model.

Observe that the model in Eq (16.133) is a direct generalization of the two-parameter model in Eq (16.19); we should therefore expect the procedure for estimating the increased number of parameters to be similar to the procedure discussed earlier. While this is true in principle, the analysis is made more tractable by using matrix methods, as we now show.

#### 16.4.1 General Least Squares

Obtaining estimates of the  $m$  parameters,  $\theta_i, i = 1, 2, \dots, m$ , from data  $(y_i; x_{1i}, x_{2i}, \dots, x_{mi})$  via the least-squares technique involves minimizing the sum-of-squares function,

$$S(\boldsymbol{\theta}) = \sum_{i=1}^n [y_i - (\theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} + \dots + \theta_m x_{mi})]^2 \quad (16.137)$$

The technique calls for taking derivatives with respect to each parameter, setting the derivative to zero and solving the resultant equations for the unknown

parameters, i.e.,

$$\frac{\partial S}{\partial \theta_0} = -2 \sum_{i=1}^n [y_i - (\theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} + \cdots + \theta_m x_{mi})] = 0 \quad (16.138)$$

and for  $1 \leq j \leq m$ ,

$$\frac{\partial S}{\partial \theta_j} = -2 \sum_{i=1}^n x_{ji} [y_i - (\theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} + \cdots + \theta_m x_{mi})] = 0 \quad (16.139)$$

These expressions rearrange to give the general linear regression normal equations:

$$\begin{aligned} \sum_{i=1}^n y_i &= \theta_0 n + \theta_1 \sum_{i=1}^n x_{1i} + \theta_2 \sum_{i=1}^n x_{2i} + \cdots + \theta_m \sum_{i=1}^n x_{mi} \\ \sum_{i=1}^n y_i x_{1i} &= \theta_0 \sum_{i=1}^n x_{1i} + \theta_1 \sum_{i=1}^n x_{1i}^2 + \theta_2 \sum_{i=1}^n x_{2i} x_{1i} + \cdots + \theta_m \sum_{i=1}^n x_{mi} x_{1i} \\ \sum_{i=1}^n y_i x_{ji} &= \theta_0 \sum_{i=1}^n x_{ji} + \theta_1 \sum_{i=1}^n x_{1i} x_{ji} + \theta_2 \sum_{i=1}^n x_{2i} x_{ji} + \cdots + \theta_m \sum_{i=1}^n x_{mi} x_{ji} \end{aligned}$$

$m+1$  linear equations to be solved simultaneously to produce the least-squares estimates for the  $m+1$  unknown parameters. Even with a modest number of parameters, such problems are best solved using matrices.

#### 16.4.2 Matrix Methods

For specific data sets,  $(y_i; x_{1i}, x_{2i}, \dots, x_{mi}); i = 1, 2, \dots, n$ , the multiple regression model equation in Eq (16.133) may be written as,

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_m \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \quad (16.140)$$

which may be consolidated into the explicit matrix form,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon} \quad (16.141)$$

where  $\mathbf{y}$  is the  $n$ -dimensional vector of response observations, with  $\mathbf{X}$  as the  $n \times m$  matrix of values of the predictor variables used to generate the  $n$  observed responses;  $\boldsymbol{\theta}$  is the  $m$ -dimensional vector of unknown parameters, and  $\boldsymbol{\epsilon}$  is the  $n$ -dimensional vector of random errors associated with the response observations. Obtaining the least-squares estimate of the parameter vector involves the same principle of minimizing the sum of squares, which, this time is given by

$$S(\boldsymbol{\theta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) \quad (16.142)$$

where the superscript,  $T$ , represents the vector or matrix transpose. Taking derivatives with respect to the parameter vector  $\boldsymbol{\theta}$  and setting the result to zero yields:

$$\frac{\partial S}{\partial \boldsymbol{\theta}} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) = 0 \quad (16.143)$$

resulting finally in the matrix form of the normal equations:

$$\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\boldsymbol{\theta} \quad (16.144)$$

(compare with series of equations shown earlier). This matrix equation is easily solved for the unknown parameter vector to produce the least squares solution:

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \quad (16.145)$$

### Properties of the Estimates

To characterize the estimates, we begin by introducing Eq (16.141) into Eq (16.145) for  $\mathbf{y}$  to obtain:

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}) \\ &= \boldsymbol{\theta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\epsilon} \end{aligned} \quad (16.146)$$

We may now use this expression to obtain the mean and variance of these estimates as follows. First, by taking expectations, we obtain:

$$\begin{aligned} E(\hat{\boldsymbol{\theta}}) &= \boldsymbol{\theta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E(\boldsymbol{\epsilon}) \\ &= \boldsymbol{\theta} \end{aligned} \quad (16.147)$$

because  $\mathbf{X}$  is known and  $E(\boldsymbol{\epsilon}) = 0$ . Thus, the least-squares estimate  $\hat{\boldsymbol{\theta}}$  as given in Eq (16.145) is unbiased for  $\boldsymbol{\theta}$ . As for the co-variance of the estimates, first, by definition,  $E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T) = \boldsymbol{\Sigma}$  is the random error covariance matrix; then from the assumption that each  $\epsilon_i$  is independent, and identically distributed, with the same variance,  $\sigma^2$ , we have that

$$\boldsymbol{\Sigma} = \sigma^2 \mathbf{I} \quad (16.148)$$

where  $\sigma^2$  is the variance associated with each random error, and  $\mathbf{I}$  is the identity matrix. As a result,

$$Var(\hat{\boldsymbol{\theta}}) = E[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T] \quad (16.149)$$

which, from Eq (16.146) becomes

$$\begin{aligned}
 Var(\hat{\theta}) &= E \left[ (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\epsilon} \boldsymbol{\epsilon}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \right] \\
 &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E(\boldsymbol{\epsilon} \boldsymbol{\epsilon}^T) \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \\
 &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \sigma^2 \mathbf{I} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \\
 &= (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2
 \end{aligned} \tag{16.150}$$

Thus, the covariance matrix of the estimates,  $\Sigma_{\hat{\theta}}$  is given by

$$\Sigma_{\hat{\theta}} = (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2 \tag{16.151}$$

As usual,  $\sigma^2$  must be estimated from data. With  $\hat{\mathbf{y}}$ , the model estimate of the response data vector  $\mathbf{y}$  now given by:

$$\hat{\mathbf{y}} = \mathbf{X} \hat{\theta} \tag{16.152}$$

the residual error vector,  $\mathbf{e}$ , is therefore defined as:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} \tag{16.153}$$

so that the residual error sum-of-squares will be given by

$$SS_E = \mathbf{e}^T \mathbf{e} \tag{16.154}$$

It can be shown that with  $p = m + 1$  as the number of estimated parameters, the mean error sum-of-squares

$$s_e = \frac{\mathbf{e}^T \mathbf{e}}{n - p} \tag{16.155}$$

is an unbiased estimate of  $\sigma$ .

Thus, following the typical normality assumption on the random error component of the regression model, we now conclude that the least-squares estimate vector,  $\hat{\theta}$ , has a multivariate normal distribution,  $MVN(\hat{\theta}, \Sigma_{\hat{\theta}})$ , with the covariance matrix as given in Eq (16.151). This fact is used to test hypotheses regarding the significance or otherwise of the parameters in precisely the same manner as before. The  $t$ -statistic arises directly from substituting data estimate  $s_e$  for  $\sigma$  in Eq (16.151).

Thus, when cast in matrix form, the multiple regression problem, is seen to be merely a higher dimensional form of the earlier simple linear regression problem; the model equations are structurally similar:

$$\begin{aligned}
 \mathbf{y} &= \mathbf{X} \theta + \epsilon \\
 y &= \theta x + \epsilon
 \end{aligned}$$

as are the least-squares solutions:

$$\begin{aligned}\hat{\theta} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \\ \hat{\theta} &= \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \\ \text{or } \hat{\theta} &= \frac{S_{xy}}{S_{xx}}\end{aligned}\quad (16.156)$$

The computations for multiple regression problems become rapidly more complex, but all the results obtained earlier for the simpler regression problem transfer virtually intact, including hypothesis tests of significance for the parameters, the values for the coefficient of determination,  $R^2$  (and its “adjusted” variant,  $R^2_{adj}$ ) for assessing the model adequacy in capturing the data information. Fortunately, these computations are routinely carried out very conveniently by computer software packages. Nevertheless, the reader is reminded that the availability of these computer programs has relieved us *only* of the computational burden; the task of *understanding* what these computations are based on remains very much an important responsibility of the practitioner.

### Residuals Analysis

The residuals in multiple linear regression are given by Eq (16.153) above. Obtaining the standardized version of residuals in this case requires the introduction of a new matrix,  $\mathbf{H}$ , the so-called “hat matrix”. If we introduce the least-squares estimate into the expression for the vector of model estimates in Eq (16.152), we obtain:

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \\ &= \mathbf{H} \mathbf{y}\end{aligned}\quad (16.157)$$

where

$$\mathbf{H} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \quad (16.158)$$

is called the “hat” matrix because it relates the actual observations vector,  $\mathbf{y}$ , to the vector of model fits,  $\hat{\mathbf{y}}$ . This matrix has some unique characteristics: for example, it is an idempotent matrix, meaning that

$$\mathbf{H} \mathbf{H} = \mathbf{H} \quad (16.159)$$

The residual vector,  $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ , may therefore be represented as:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H}) \mathbf{y} \quad (16.160)$$

(The matrix  $(\mathbf{I} - \mathbf{H})$  is also idempotent). If  $h_{ii}$  represents the diagonal elements of  $\mathbf{H}$ , the standardized residuals are obtained for multiple regression problems

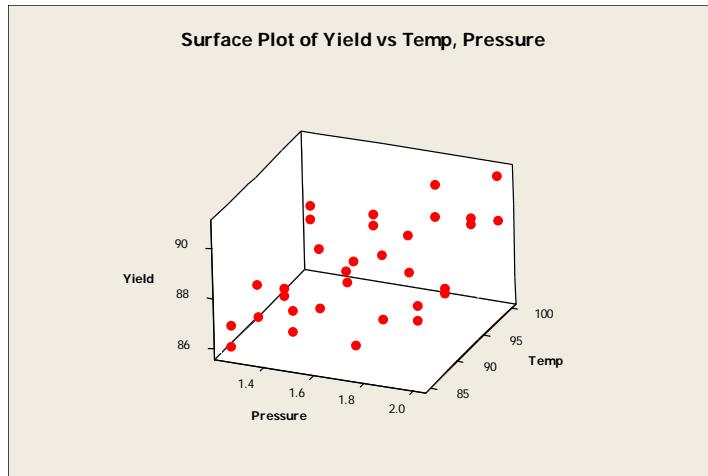


FIGURE 16.10: Catalytic process yield data of Table 16.7

as:

$$e_i^* = \frac{e_i}{s_e \sqrt{1 - h_{ii}}} \quad (16.161)$$

These standardized residuals are the exact equivalents of the ones shown in Eq (16.116) for the simple linear regression case.

The next example illustrates an application of these results.

#### **Example 16.8: QUANTIFYING TEMPERATURE AND PRESSURE EFFECTS ON YIELD**

In an attempt to quantify the effect of temperature and pressure on the yield obtained from a laboratory scale catalytic process, the data shown in Table 16.7 was obtained from a series of designed experiments where temperature was varied over a relatively narrow range, from  $85^\circ C$  to  $100^\circ C$ , and pressure from 1.25 atmospheres to 2 atmospheres. If Yield is  $y$ , Temperature is  $x_1$  and Pressure,  $x_2$ , obtain a regression model of the type in Eq (16.135) and evaluate the model fit.

#### **Solution:**

A 3-D scatter plot of the data is shown in Fig 16.10, where it appears as if the data truly fall on a plane.

The results from an analysis carried out using MINITAB is as follows:

#### **Regression Analysis: Yield versus Temp, Pressure**

The regression equation is

$$\underline{75.9 + 0.0757 \text{ Temp} + 3.21 \text{ Pressure}}$$

| Predictor | Coef    | SE Coef | T     | P     |
|-----------|---------|---------|-------|-------|
| Constant  | 75.866  | 2.924   | 25.95 | 0.000 |
| Temp      | 0.07574 | 0.02977 | 2.54  | 0.017 |
| Pressure  | 3.2120  | 0.5955  | 5.39  | 0.000 |

**TABLE 16.7:** Laboratory experimental data on Yield obtained from a catalytic process at various temperatures and pressures

| Yield (%) | Temp (°C) | Pressure (Atm) |
|-----------|-----------|----------------|
| 86.8284   | 85        | 1.25           |
| 87.4136   | 90        | 1.25           |
| 86.2096   | 95        | 1.25           |
| 87.8780   | 100       | 1.25           |
| 86.9892   | 85        | 1.50           |
| 86.8632   | 90        | 1.50           |
| 86.8389   | 95        | 1.50           |
| 88.0432   | 100       | 1.50           |
| 86.8420   | 85        | 1.75           |
| 89.3775   | 90        | 1.75           |
| 87.6432   | 95        | 1.75           |
| 90.0723   | 100       | 1.75           |
| 88.8353   | 85        | 2.00           |
| 88.4265   | 90        | 2.00           |
| 90.1930   | 95        | 2.00           |
| 89.0571   | 100       | 2.00           |
| 85.9974   | 85        | 1.25           |
| 86.1209   | 90        | 1.25           |
| 85.8819   | 95        | 1.25           |
| 88.4381   | 100       | 1.25           |
| 87.8307   | 85        | 1.50           |
| 89.2073   | 90        | 1.50           |
| 87.2984   | 95        | 1.50           |
| 88.5071   | 100       | 1.50           |
| 90.1824   | 85        | 1.75           |
| 86.8078   | 90        | 1.75           |
| 89.1249   | 95        | 1.75           |
| 88.7684   | 100       | 1.75           |
| 88.2137   | 85        | 2.00           |
| 88.2571   | 90        | 2.00           |
| 89.9551   | 95        | 2.00           |
| 90.8301   | 100       | 2.00           |

$S = 0.941538$  R-Sq = 55.1 % R-Sq(adj) = 52.0%

Thus, the fitted regression line equation is, in this case

$$\hat{y} = 75.9 + 0.0757x_1 + 3.21x_2 \quad (16.162)$$

The  $p$ -values associated with all the parameters indicate significance; the estimate of the error standard deviation is as shown (0.94) with the  $R^2$  and  $R_{adj}^2$  values indicating that the model explanation of the variation in the data is only modest.

The fitted plane represented by Eq (16.162) and the standardized residual errors are shown in Fig 16.11. There is nothing unusual about the residuals but the relatively modest values of  $R^2$  and  $R_{adj}^2$  seem to suggest that the true model might be somewhat more complicated than the one we have postulated and fitted here; it could also mean that there is significant noise associated with the measurement, or both.

### 16.4.3 Some Important Special Cases

#### Weighted Least Squares

For a wide variety of reasons, ranging from  $x_i$  variables with values that are orders of magnitude apart (e.g., if  $x_1$  is temperature in the 100's of degrees, while  $x_2$  is mole fraction, naturally scaled between 0 and 1), to measurement errors with non-constant variance-covariance structures, it is often necessary to modify the basic multiple linear regression problem by placing more or less weight on different observations. Under these circumstances, the regression model equation in Eq (16.141) is modified to:

$$\mathbf{W}\mathbf{y} = \mathbf{W}\mathbf{X}\boldsymbol{\theta} + \mathbf{W}\boldsymbol{\epsilon} \quad (16.163)$$

where  $\mathbf{W}$  is an appropriately chosen ( $n \times n$ ) weighting matrix. Note that the pre-multiplication in Eq (16.163) has not changed the model itself; it merely allows a re-scaling of the  $\mathbf{X}$  matrix and/or the error vector. However, the introduction of the weights does affect the solution to the least-squares optimization problem. It can be shown that in this case, the sum-of-squares

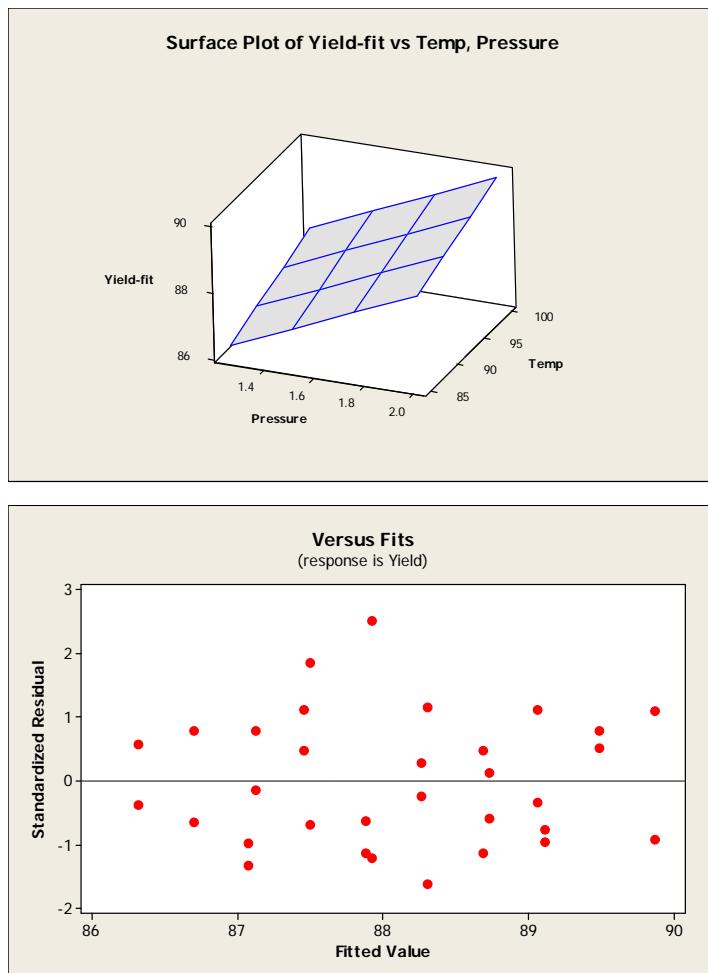
$$S(\boldsymbol{\theta}) = (\mathbf{W}\mathbf{y} - \mathbf{W}\mathbf{X}\boldsymbol{\theta})^T(\mathbf{W}\mathbf{y} - \mathbf{W}\mathbf{X}\boldsymbol{\theta}) \quad (16.164)$$

is minimized by

$$\hat{\boldsymbol{\theta}}_{WLS} = \left( \mathbf{X}^T \mathbf{W}^T \mathbf{W} \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{W}^T \mathbf{W} \mathbf{y} \quad (16.165)$$

This is known as the weighted least-squares (WLS) estimate, and it is easy to establish that regardless of the choice of  $\mathbf{W}$ ,

$$E(\hat{\boldsymbol{\theta}}_{WLS}) = \boldsymbol{\theta} \quad (16.166)$$



**FIGURE 16.11:** Catalytic process yield data of Table 16.1. Top: Fitted plane of Yield as a function of Temperature and Pressure; Bottom: standardized residuals versus fitted value  $\hat{y}_i$ . Nothing appears unusual about these residuals.

so that this is also an unbiased estimate of  $\boldsymbol{\theta}$ .

In cases where the motivation for introducing weights arises from the structure of the error covariance matrix,  $\Sigma$ , it is recommended that  $\mathbf{W}$  be chosen such that

$$\mathbf{W}^T \mathbf{W} = \Sigma^{-1} \quad (16.167)$$

Under these circumstances, the covariance matrix of the WLS estimate can be shown to be given by:

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_{WLS}} = (\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} \quad (16.168)$$

making it comparable to Eq (16.151). All regression packages provide an option for carrying out WLS instead of ordinary least squares.

### Constrained Least Squares

Occasionally, one encounters regression problems in engineering where the model parameters are subject to a set of linear equality constraints. For example, in a blending problem where the unknown parameters,  $\theta_1, \theta_2$ , and  $\theta_3$  to be estimated from experimental data, are the mole fractions of the three component materials, clearly

$$\theta_1 + \theta_2 + \theta_3 = 1 \quad (16.169)$$

In general such linear constraints are of the form:

$$\mathbf{L}\boldsymbol{\theta} = \mathbf{v} \quad (16.170)$$

When subject to such constraints, obtaining the least squares estimate of the parameters in the model Eq (16.141) now requires attaching these constraint equations to the original problem of minimizing the sum of squares function  $S(\boldsymbol{\theta})$ . It can be shown that when the standard tools of Lagrange multipliers are used to solve this constrained optimization problem, the solution is:

$$\hat{\boldsymbol{\theta}}_{CLS} = \hat{\boldsymbol{\theta}} + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{L}^T \left[ \mathbf{L} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{L}^T \right]^{-1} (\mathbf{v} - \mathbf{L}\hat{\boldsymbol{\theta}}) \quad (16.171)$$

the constrained least squares (CLS) estimate, where  $\hat{\boldsymbol{\theta}}$  is the normal, unconstrained least-squares estimate in Eq (16.145).

If we define a “gain” matrix:

$$\boldsymbol{\Gamma} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{L}^T \left[ \mathbf{L} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{L}^T \right]^{-1} \quad (16.172)$$

then, Eq (16.171) may be rearranged to:

$$\hat{\boldsymbol{\theta}}_{CLS} = \hat{\boldsymbol{\theta}} + \boldsymbol{\Gamma}(\mathbf{v} - \mathbf{L}\hat{\boldsymbol{\theta}}) \quad (16.173)$$

$$\text{or } \hat{\boldsymbol{\theta}}_{CLS} = \boldsymbol{\Gamma}\mathbf{v} + (\mathbf{I} - \boldsymbol{\Gamma}\mathbf{L})\hat{\boldsymbol{\theta}} \quad (16.174)$$

where the former (as in Eq (16.171)) emphasizes how the constraints provide a correction to the unconstrained estimate, and the latter emphasizes that  $\hat{\theta}_{CLS}$  provides a compromise between unconstrained estimates and the constraints.

### Ridge Regression

The ordinary least squares estimate,  $\hat{\theta}$ , given in Eq (16.145), will be unacceptable for “ill-conditioned” problems for which  $\mathbf{X}^T \mathbf{X}$  is nearly singular typically because the determinant,  $|\mathbf{X}^T \mathbf{X}| \approx 0$ . This will occur, for example, when there is near-linear dependence in some of the predictor variables,  $x_i$ , or when some of the  $x_i$  variables are orders of magnitude different from others, and the problem has not been re-scaled accordingly. The problem created by ill-conditioning manifests in the form of overly inflated values for the elements of the matrix inverse  $(\mathbf{X}^T \mathbf{X})^{-1}$  as a result of the vanishingly small determinant. Consequently, the norm of the estimate vector,  $\hat{\theta}$ , will be too large, and the uncertainty associated with the estimates (see Eq 16.151) will be unacceptably large.

One solution is to augment the original model equation as follows:

$$\begin{bmatrix} \mathbf{y} \\ \vdots \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{X} \\ \vdots \\ k\mathbf{I} \end{bmatrix} \boldsymbol{\theta} + \boldsymbol{\epsilon} \quad (16.175)$$

or,

$$\mathbf{y}^* = \mathbf{X}^* \boldsymbol{\theta} + \boldsymbol{\epsilon} \quad (16.176)$$

where  $\mathbf{0}$  is an  $m$ -dimensional vector of zeros,  $k$  is a constant, and  $\mathbf{I}$  is the identity matrix. Instead of minimizing the original sum of squares function, minimizing the sum of squares based on the augmented equation (16.176) results in the so-called ridge regression estimate:

$$\hat{\boldsymbol{\theta}}_{RR} = (\mathbf{X}^T \mathbf{X} + k^2 \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y} \quad (16.177)$$

As is evident from Eq (16.175), the purpose of the augmentation is to use the constant  $k$  to force the estimated parameters to be close to 0, preventing over-inflation. As the value chosen for  $k$  increases, the conditioning of the matrix  $(\mathbf{X}^T \mathbf{X} + k^2 \mathbf{I})$  improves, reducing the otherwise inflated estimates  $\hat{\boldsymbol{\theta}}$ . However, this improvement is achieved at the expense of introducing bias into the resulting estimate vector. Still, even though  $\hat{\boldsymbol{\theta}}_{RR}$  is biased, its variance is much better than that of the original  $\hat{\boldsymbol{\theta}}$ . (See Hoerl, (1962)<sup>1</sup>, Hoerl and

---

<sup>1</sup>Hoerl, A.E. (1962). “Application of ridge analysis to regression problems,” *Chem. Eng. Prog.* 55, 54–59.

Kennard, (1970a)<sup>2</sup>, (1970b)<sup>3</sup>). Selecting an appropriate value of  $k$  is an art. (See Marquardt, (1970)<sup>4</sup>).

#### 16.4.4 Recursive Least Squares

##### Problem Formulation

A case often arises in engineering where the experimental data used to estimate parameters in the model in Eq (16.141) are available sequentially. After accumulating a set of  $n$  observations,  $y_i; i = 1, 2, \dots, n$ , and, subsequently using this  $n$ -dimensional data vector,  $\mathbf{y}_n$ , to obtain the parameter estimates as:

$$\hat{\boldsymbol{\theta}}_n = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}_n \quad (16.178)$$

suppose that a new, single observation  $y_{n+1}$  then becomes available. This new data can be combined with past information to obtain an updated estimate that reflects the additional information about the parameter contained in the data, information represented by:

$$y_{n+1} = \mathbf{x}_{n+1}^T \boldsymbol{\theta} + \epsilon_{n+1} \quad (16.179)$$

where  $\mathbf{x}_{n+1}^T$  is the  $m$ -dimensional vector of the independent predictor variables used to generate this new  $(n + 1)^{th}$  observation, and  $\epsilon_{n+1}$  is the associated random error component.

In principle, we can append this new information to the old to obtain:

$$\begin{bmatrix} \mathbf{y}_n \\ \vdots \\ y_{n+1} \end{bmatrix} = \begin{bmatrix} \mathbf{X} \\ \vdots \\ \mathbf{x}_{n+1}^T \end{bmatrix} \boldsymbol{\theta} + \begin{bmatrix} \boldsymbol{\epsilon}_n \\ \vdots \\ \epsilon_{n+1} \end{bmatrix} \quad (16.180)$$

or, more compactly,

$$\mathbf{y}_{n+1} = \mathbf{X}_{n+1} \boldsymbol{\theta} + \boldsymbol{\epsilon}_{n+1} \quad (16.181)$$

so that the new  $\mathbf{X}$  matrix,  $\mathbf{X}_{n+1}$ , is now an  $(n + 1) \times m$  matrix, with the data vector  $\mathbf{y}_{n+1}$  now of dimension  $(n + 1)$ . From here, we can use these new matrices and vectors to obtain the new least-squares estimate directly, as before, to give:

$$\hat{\boldsymbol{\theta}}_{n+1} = (\mathbf{X}_{n+1}^T \mathbf{X}_{n+1})^{-1} \mathbf{X}_{n+1}^T \mathbf{y}_{n+1} \quad (16.182)$$

Again, in principle, we can repeat this exercise each time a new observation becomes available. However, such a strategy requires that we recompute the

---

<sup>2</sup>Hoerl A.E., and R.W. Kennard, (1970). "Ridge regression. Biased estimation for nonorthogonal problems," *Technometrics*, 12, 55–67.

<sup>3</sup>Hoerl A.E., and R.W. Kennard, (1970). "Ridge regression. Applications to nonorthogonal problems," *Technometrics*, 12, 69–82.

<sup>4</sup>Marquardt, D.W., (1970). "Generalized inverses, Ridge regression, Biased linear estimation, and Nonlinear estimation," *Technometrics*, 12, 591–612.

estimates from scratch every time as if for the first time. While it is true that the indicated computational burden is routinely borne nowadays by computers, the fact that the information is available recursively raises a fundamental question: *instead of having to recompute the estimate  $\hat{\theta}_{n+1}$  all over again as in Eq (16.182) every time new information is available, is it possible to determine it by judiciously updating  $\hat{\theta}_n$  directly with the new information?* The answer is provided by the recursive least-squares technique whereby  $\hat{\theta}_{n+1}$  is obtained recursively as a function of  $\hat{\theta}_n$ .

### Recursive Least Squares Estimation

We begin by obtaining the least-squares estimate,  $\hat{\theta}_{n+1}$ , directly from the partitioned matrices in Eq (16.180), giving the result

$$\hat{\theta}_{n+1} = \left[ \mathbf{X}^T \mathbf{X} + \mathbf{x}_{n+1} \mathbf{x}_{n+1}^T \right]^{-1} \left[ \mathbf{X}^T \mathbf{y}_n + \mathbf{x}_{n+1} y_{n+1} \right] \quad (16.183)$$

Now, let us define

$$\mathbf{P}_{n+1} = \left[ \mathbf{X}^T \mathbf{X} + \mathbf{x}_{n+1} \mathbf{x}_{n+1}^T \right]^{-1} \quad (16.184)$$

so that by analogy with Eq (16.178),

$$\mathbf{P}_n = \left( \mathbf{X}^T \mathbf{X} \right)^{-1}; \quad (16.185)$$

then,

$$\mathbf{P}_{n+1}^{-1} = \mathbf{X}^T \mathbf{X} + \mathbf{x}_{n+1} \mathbf{x}_{n+1}^T$$

From here, on the one hand, by simple rearrangement,

$$\mathbf{X}^T \mathbf{X} = \mathbf{P}_{n+1}^{-1} - \mathbf{x}_{n+1} \mathbf{x}_{n+1}^T \quad (16.186)$$

and on the other, using Eqn (16.185),

$$\mathbf{P}_{n+1}^{-1} = \mathbf{P}_n^{-1} + \mathbf{x}_{n+1} \mathbf{x}_{n+1}^T \quad (16.187)$$

Now, as a result of Eq (16.184), the least-squares estimate in Eq (16.183) may be written as:

$$\begin{aligned} \hat{\theta}_{n+1} &= \mathbf{P}_{n+1} \left[ \mathbf{X}^T \mathbf{y}_n + \mathbf{x}_{n+1} y_{n+1} \right] \\ &= \mathbf{P}_{n+1} \mathbf{X}^T \mathbf{y}_n + \mathbf{P}_{n+1} \mathbf{x}_{n+1} y_{n+1} \end{aligned} \quad (16.188)$$

Returning briefly to Eq (16.178) from which

$$(\mathbf{X}^T \mathbf{X}) \hat{\theta}_n = \mathbf{X}^T \mathbf{y}_n,$$

upon introducing Eq (16.186), we obtain

$$(\mathbf{P}_{n+1}^{-1} - \mathbf{x}_{n+1} \mathbf{x}_{n+1}^T) \hat{\theta}_n = \mathbf{X}^T \mathbf{y}_n \quad (16.189)$$

Introducing this into Eq (16.188) yields

$$\begin{aligned}\hat{\theta}_{n+1} &= \mathbf{P}_{n+1} (\mathbf{P}_{n+1}^{-1} - \mathbf{x}_{n+1} \mathbf{x}_{n+1}^T) \hat{\theta}_n + \mathbf{P}_{n+1} \mathbf{x}_{n+1} y_{n+1} \\ \text{or, } \hat{\theta}_{n+1} &= \hat{\theta}_n + \mathbf{P}_{n+1} \mathbf{x}_{n+1} \left( y_{n+1} - \mathbf{x}_{n+1}^T \hat{\theta}_n \right)\end{aligned}\quad (16.190)$$

This last equation is the required recursive expression for determining  $\hat{\theta}_{n+1}$  from  $\hat{\theta}_n$  and new response data,  $y_{n+1}$ , generated with the new values  $\mathbf{x}_{n+1}^T$  for the predictor variables. The “gain matrix,”  $\mathbf{P}_{n+1}$ , is itself generated recursively from Eq (16.187). And now, several points worth noting:

1. The matrix  $\mathbf{P}_n$  is related to the covariance matrix of the estimate,  $\hat{\theta}_n$ , (see Eq (16.151)), so that Eq (16.187) represents the recursive evolution of the covariance of the estimates as  $n$  increases;
2. The term in parentheses in Eq (16.190) resembles a correction term, the discrepancy between the actual observed response,  $y_{n+1}$ , and an *a priori* value predicted for it (before the new data is available) using the previous estimate,  $\hat{\theta}_n$ , and the new predictor variables,  $\mathbf{x}_{n+1}^T$ ;
3. This recursive procedure allows us to begin with an initial estimate,  $\hat{\theta}_0$ , along with a corresponding (scaled) covariance matrix,  $\mathbf{P}_0$ , and proceed recursively to estimate the true value of the parameters one data point at a time, using Eq (16.187) first to obtain an updated covariance matrix, and Eq (16.190) to update the parameter estimate;
4. Readers familiar with Kalman filtering in dynamical systems theory will immediately recognize the structural similarity between the combination of Eqs (16.187) and (16.190) and the discrete Kalman filter.

## 16.5 Polynomial Regression

### 16.5.1 General Considerations

A special case of multiple linear regression occurs when, in Eq (16.133), the response  $Y$  depends on powers of a single predictor variable,  $x$ , i.e.,

$$Y = \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_m x^m + \epsilon \quad (16.191)$$

in which case  $x_i = x^i$ . An example was given earlier in Eq (16.135), with  $Y$  as a quadratic function of  $x$ .

This class of regression models is important because in engineering, many unknown functional relationships,  $y(x)$ , can be approximated by polynomials. Because Eq (16.191) is a special case of Eq (16.133), all the results obtained

earlier for the more general problem transfer directly, and there is not much to add for this restricted class of problems. However, in terms of practical application, there are some peculiarities unique to polynomial regression analysis.

In many practical problems, the starting point in polynomial regression is often a low order linear model; when residual analysis indicates that the simple model is inadequate, the model complexity is then increased, typically by adding the next higher power of  $x$ , until the model is deemed "adequate." But one must be careful: fitting an  $m^{th}$  order polynomial to  $m+1$  data points (e.g. fitting a straight line to 2 points) will produce a perfect  $R^2 = 1$  but the parameter estimates will be unreliable. The primary pitfall to avoid in such an exercise is therefore "overfitting," whereby the polynomial model is of an order higher than can be realistically supported by the data. Under such circumstances, the improvement in  $R^2$  must be cross-checked against the corresponding  $R_{adj}^2$  value.

The next examples illustrate the application of polynomial regression.

**Example 16.9: BOILING POINT OF HYDROCARBONS:  
REVISITED**

In Example 16.6, a linear two-parameter model was postulated for the relationship between the number of carbon atoms in the hydrocarbon compounds listed in Table 16.1 and the respective boiling points. Upon evaluation, however, the model was found to be inadequate; specifically, the residuals indicated the potential for a "left over" quadratic component. Postulate the following quadratic model,

$$Y = \theta_0 + \theta_1 x + \theta_2 x^2 + \epsilon \quad (16.192)$$

and evaluate a least-squares fit of this model to the data. Compare this model fit to the simple linear model obtained in Example 16.6.

**Solution:**

Once more, when we use MINITAB for this problem, we obtain the following results:

**Regression Analysis: Boiling Point versus n, n2**

The regression equation is

| Predictor | Coef     | SE Coef | T      | P     |
|-----------|----------|---------|--------|-------|
| Constant  | -218.143 | 8.842   | -24.67 | 0.000 |
| n         | 66.667   | 4.508   | 14.79  | 0.000 |
| n2        | -3.0238  | 0.4889  | -6.18  | 0.002 |

S = 6.33734 R-Sq = 99.7% R-Sq(adj) = 99.6%

Thus, the fitted quadratic regression line equation is

$$\hat{y} = -218.14 + 66.67x - 3.02x^2 \quad (16.193)$$

where, as before,  $y$  is the hydrocarbon compound boiling point, and the number of carbon atoms it contains is  $x$ .

Note how the estimates for  $\theta_0$  and  $\theta_1$  are now different from the respective values obtained when these were the only two parameters in the model. This is a natural consequence of adding a new component to the model; the responsibility for capturing the variability in the data is now being shared by three parameters instead of 2, and the best estimates of the model parameter set will change accordingly.

Before inspecting the model fit and the residuals, we note first that the three parameters in this case also are all significantly different from zero (the  $p$ -values are zero for the constant term and the linear term and 0.002 for the quadratic coefficient). As expected, there is an improvement in  $R^2$  for this more complicated model (99.7% versus 97.4% for the simpler linear model); furthermore, this improvement was also accompanied by a commensurate improvement in  $R_{adj}^2$  (99.6% versus 97.0% for the simpler model). Thus, the improved model performance was not achieved at the expense of overfitting, indicating that the added quadratic term is truly warranted. The error standard deviation also shows an almost three-fold improvement from  $S = 17.0142$  for the linear model to  $S = 6.33734$ , again indicating that more of the variability in the data has been captured by the more complicated model.

The model fit to the data, shown in the top panel of Fig 16.12, indicates a much-improved fit, compared with the one in the top panel of Fig 16.9. This is also consistent with everything we have noted so far. However, the normalized residuals versus fitted values, plotted in the bottom panel of Fig 16.12 shows that there is still some “left over” structure, the improved fit notwithstanding. The implication is that perhaps an additional cubic term might be necessary to capture the remaining structural information still visible in the residual plot. This suggests further revising the model as follows:

$$Y = \theta_0 + \theta_1x + \theta_2x^2 + \theta_3x^3 + \epsilon \quad (16.194)$$

The problem of fitting an adequate regression model to the data in Table 16.1 concludes with this next example.

**Example 16.10: BOILING POINT OF HYDROCARBONS:  
PART III**

As a follow up to the analysis in the last example, fit the cubic equation

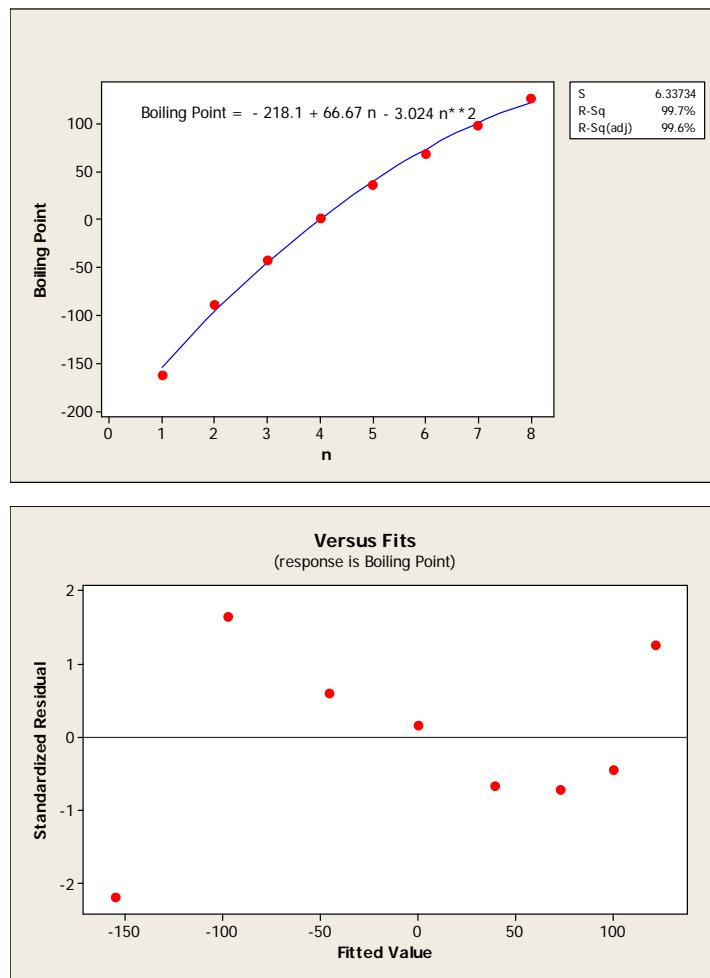
$$Y = \theta_0 + \theta_1x + \theta_2x^2 + \theta_3x^3 + \epsilon \quad (16.195)$$

to the data in Table 16.1, evaluate the model fit and compare it to the fit obtained in Example 16.9.

**Solution:**

This time around, the MINITAB results are as follows:

**Regression Analysis: Boiling Point versus n, n2, n3**  
The regression equation is



**FIGURE 16.12:** Modeling the dependence of the boiling points (BP) of hydrocarbon compounds in Table 16.1 on the number of carbon atoms in the compound: Top: Fitted quadratic curve of BP versus  $n$ , the number of carbon atoms; Bottom: standardized residuals versus fitted value  $\hat{y}_i$ . Despite the good fit, the visible systematic structure still “left over” in the residuals suggests adding one more term to the model.

| Boiling Point = - 244 + 93.2 n - 9.98 n <sup>2</sup> + 0.515 n <sup>3</sup> |          |         |        |       |
|---|----------|---------|--------|-------|
| Predictor   | Coef     | SE Coef | T      | P     |
| Constant  | -243.643 | 8.095   | -30.10 | 0.000 |
| n   | 93.197   | 7.325   | 12.72  | 0.000 |
| n <sup>2</sup>  | -9.978   | 1.837   | -5.43  | 0.006 |
| n <sup>3</sup>  | 0.5152   | 0.1348  | 3.82   | 0.019 |

S = 3.28531 R-Sq = 99.9% R-Sq(adj) = 99.9%

The fitted cubic regression equation is

$$\hat{y} = -243.64 + 93.20x - 9.98x^2 + 0.515x^3 \quad (16.196)$$

Note that the estimates for  $\theta_0$  and  $\theta_1$  have changed once again, as has the estimate for  $\theta_2$ . Again, this is a natural consequence of adding the new parameter,  $\theta_3$ , to the model.

As a result of the p-values, we conclude once again, that all the four parameters in this model are significantly different from zero; the  $R^2$  and  $R_{adj}^2$  values are virtually perfect and identical, indicating that the expenditure of four parameters in this model is justified.

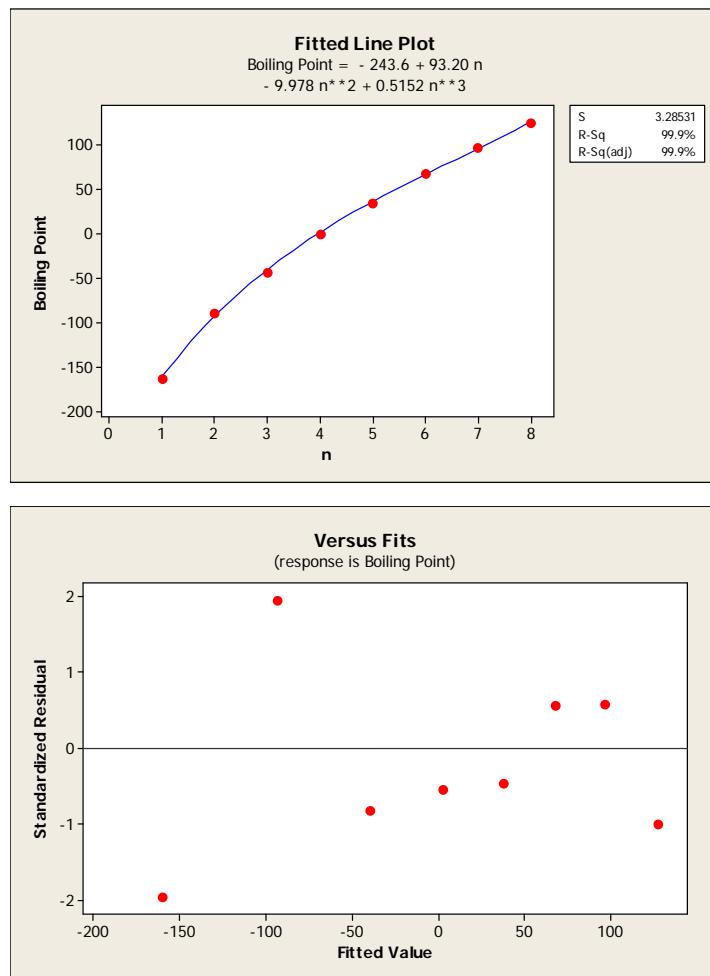
The error standard deviation has improved further by a factor of almost 2 (from  $S = 6.33734$  for the quadratic model to  $S = 3.28531$  for this cubic model) and the model fit to the data shows this improvement graphically in the top panel of Fig 16.13. This time, the residual plot in the bottom panel of Fig 16.13 shows no significant “left over” structure. Therefore, in light of all the factors considered above, we conclude that the cubic fit in Eq 16.196 appears to provide an adequate fit to the data; and that this has been achieved without the expenditure of an excessive number of parameters.

A final word: while polynomial models may provide adequate representations of data (as in the last example), this should not be confused with a fundamental scientific *explanation* of the underlying relationship between  $x$  and  $Y$ . Also, it is not advisable to extrapolate the model prediction outside of the range covered by the data used to fit the model. For example, the model in this last example should not be used to predict the boiling point of hydrocarbons with carbon atoms  $\geq 9$ .

### 16.5.2 Orthogonal Polynomial Regression

Let  $p_i(x); i = 0, 1, 2, \dots, m$ , be a sequence of  $i^{th}$  order polynomials in the single independent variable,  $x$ , defined on an interval  $[x_L, x_R]$  on the real line. For our purposes here, these polynomials take on *discrete* values  $p_i(x_k)$  at equally spaced values  $x_k; k = 1, 2, \dots, n$ , in the noted interval. The sequence of polynomials constitutes an orthogonal set if the following conditions hold:

$$\sum_{k=1}^n p_i(x_k) p_j(x_k) = \begin{cases} \psi_i^2 & i = j; \\ 0 & i \neq j \end{cases} \quad (16.197)$$



**FIGURE 16.13:** Modeling the dependence of the boiling points (BP) of hydrocarbon compounds in Table 16.1 on the number of carbon atoms in the compound: Top: Fitted cubic curve of BP versus  $n$ , the number of carbon atoms; Bottom: standardized residuals versus fitted value  $\hat{y}_i$ . There appears to be little or no systematic structure left in the residuals, suggesting that the cubic model provides an adequate description of the data.

### An Example: Gram Polynomials

Without loss of generality, let the independent variable  $x$  be defined in the interval  $[-1, 1]$  (for variables defined on  $[x_l, x_R]$ , a simple scaling transformation is all that is required to obtain a corresponding variable defined instead on  $[-1, 1]$ ); furthermore, let this interval be divided into  $n$  equal discrete intervals,  $k = 1, 2, \dots, n$ , to provide  $n$  values of  $x$  at  $x_1, x_2, \dots, x_n$ , such that  $x_1 = -1$ ,  $x_n = 1$ , and in general,

$$x_k = \frac{2(k-1)}{n-1} - 1 \quad (16.198)$$

The set of Gram polynomials defined on  $[-1, 1]$  for  $x_k$  as given above is,

$$\begin{aligned} p_0(x_k) &= 1 \\ p_1(x_k) &= x_k \\ p_2(x_k) &= x_k^2 - \frac{(n+1)}{3(n-1)} \\ p_3(x_k) &= x_k^3 - \left[ \frac{(3n^2-7)}{5(n-1)^2} \right] x_k \\ &\vdots & \vdots \\ p_{\ell+1}(x) &= x_k p_\ell(x_k) - \left[ \frac{\ell^2(n^2-\ell^2)}{(4\ell^2-1)(n-1)^2} \right] p_{\ell-1}(x_k) \end{aligned} \quad (16.199)$$

where each polynomial in the set is generated from the “recurrence relation” in Eq (16.199), given the initial two,  $p_0(x_k) = 1$  and  $p_1(x_k) = x_k$ .

#### **Example 16.11: ORTHOGONALITY OF GRAM POLYNOMIALS**

Obtain the first four Gram polynomials determined at  $n = 5$  equally spaced values,  $x_k$ , of the independent variable,  $x$ , on the interval  $-1 \leq x \leq 1$ . Show that these polynomials are mutually orthogonal.

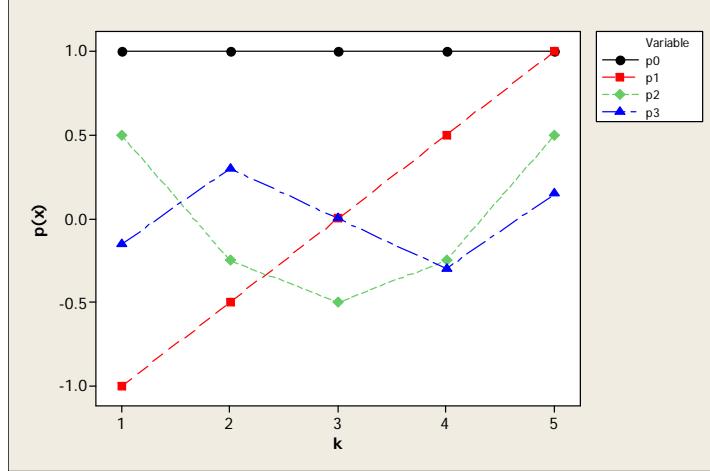
#### **Solution:**

First, from Eq (16.198), the values  $x_k$  at which the polynomials are to be determined in the interval  $1 \leq x \leq 1$  are:

$$x_1 = -1; x_2 = -0.5; x_3 = 0; x_4 = 0.5; x_5 = 1.$$

Next, let the 5-dimensional vector,  $\mathbf{p}_i; i = 0, 1, 2, 3$ , represent the values of the polynomial  $p_i(x_k)$  determined at these five  $x_k$  values: i.e.,

$$\mathbf{p}_i = \begin{bmatrix} p_i(x_1) \\ p_i(x_2) \\ \vdots \\ p_i(x_5) \end{bmatrix} \quad (16.200)$$



**FIGURE 16.14:** Gram polynomials evaluated at 5 discrete points  $k = 1, 2, 3, 4, 5$ ;  $p_0$  is the constant;  $p_1$ , the straight line;  $p_2$ , the quadratic and  $p_3$ , the cubic

Then, from the expressions given above, we obtain, using  $n = 5$ , that:

$$\mathbf{p}_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}; \mathbf{p}_1 = \begin{bmatrix} -1 \\ -0.5 \\ 0 \\ 0.5 \\ 1 \end{bmatrix}; \mathbf{p}_2 = \begin{bmatrix} 0.50 \\ -0.25 \\ -0.50 \\ -0.25 \\ 0.50 \end{bmatrix}; \mathbf{p}_3 = \begin{bmatrix} -0.15 \\ 0.30 \\ 0.00 \\ -0.30 \\ 0.15 \end{bmatrix}$$

A plot of these values is shown in Fig 16.14 where we see that  $p_0(x_k)$  is a constant,  $p_1(x_k)$  is a straight line,  $p_2(x_k)$  is a quadratic and  $p_3(x_k)$  is a cubic, each one evaluated at the indicated discrete points.

To establish orthogonality, we compute inner products,  $\mathbf{p}_i^T \mathbf{p}_j$ , for all combinations of  $i \neq j$ . First, we note that  $\mathbf{p}_0^T \mathbf{p}_j$  is simply a sum of all the elements in each vector, which is uniformly zero in all cases, i.e.,

$$\mathbf{p}_0^T \mathbf{p}_j = \sum_{k=1}^5 p_j(x_k) = 0 \quad (16.201)$$

Next, we obtain:

$$\mathbf{p}_1^T \mathbf{p}_2 = \sum_{k=1}^5 p_1(x_k) p_2(x_k) = -0.500 + 0.125 + 0.000 - 0.125 + 0.500 = 0$$

$$\mathbf{p}_1^T \mathbf{p}_3 = \sum_{k=1}^5 p_1(x_k) p_3(x_k) = 0.15 - 0.15 + 0.00 - 0.15 + 0.15 = 0$$

$$\mathbf{p}_2^T \mathbf{p}_3 = \sum_{k=1}^5 p_2(x_k) p_3(x_k) = -0.075 - 0.075 + 0.000 + 0.075 + 0.075 = 0$$

For completeness, the sums of squares,  $\psi_i^2$ , are obtained below (note monotonic decrease):

$$\psi_0^2 = \mathbf{p}_0^T \mathbf{p}_0 = \sum_{k=1}^5 p_0(x_k) p_0(x_k) = 5$$

$$\psi_1^2 = \mathbf{p}_1^T \mathbf{p}_1 = \sum_{k=1}^5 p_1(x_k) p_1(x_k) = 2.5$$

$$\psi_2^2 = \mathbf{p}_2^T \mathbf{p}_2 = \sum_{k=1}^5 p_2(x_k) p_2(x_k) = 0.875$$

$$\psi_3^2 = \mathbf{p}_3^T \mathbf{p}_3 = \sum_{k=1}^5 p_3(x_k) p_3(x_k) = 0.225$$

### Application in Regression

Among many attractive properties possessed by orthogonal polynomials, the following is the most relevant to the current discussion:

*Orthogonal Basis Function Expansion:* Any  $m^{th}$  order polynomial,  $U(x)$ , can be expanded in terms of an orthogonal polynomial set,  $p_0(x), p_1(x), \dots, p_m(x)$ , as the basis functions, i.e.,

$$U(x) = \sum_{i=0}^m \alpha_i p_i(x) \quad (16.202)$$

This result has some significant implications for polynomial regression involving the single independent variable,  $x$ . Observe that as a consequence of this result, the original  $m^{th}$  order polynomial regression model in Eq (16.191) can be rewritten as

$$Y(x) = \alpha_0 p_0(x) + \alpha_1 p_1(x) + \alpha_2 p_2(x) + \dots + \alpha_m p_m(x) + \epsilon \quad (16.203)$$

where we note that, given any specific set of orthogonal polynomial basis, the one-to-one relationship between the original parameters,  $\theta_i$ , and the new set,  $\alpha_i$ , is easily determined. Regression analysis is now concerned with estimating the new set of parameters,  $\alpha_i$ , instead of the old set,  $\theta_i$ , a task that is rendered dramatically easier because of the orthogonality of the basis set,  $p_i(x)$ , as we now show.

Suppose that the data  $y_i; i = 1, 2, \dots, n$ , have been acquired using equally spaced values  $x_k; k = 1, 2, \dots, n$ , in the range  $[x_L, x_R]$  over which the orthogonal polynomial set,  $p_i(x)$ , is defined. In this case, from Eq (16.203), we will

have:

$$\begin{aligned} y(x_1) &= \alpha_0 p_0(x_1) + \alpha_1 p_1(x_1) + \alpha_2 p_2(x_1) + \cdots + \alpha_m p_m(x_1) + \epsilon_1 \\ y(x_2) &= \alpha_0 p_0(x_2) + \alpha_1 p_1(x_2) + \alpha_2 p_2(x_2) + \cdots + \alpha_m p_m(x_2) + \epsilon_2 \\ &\vdots && \vdots \\ y(x_n) &= \alpha_0 p_0(x_n) + \alpha_1 p_1(x_n) + \alpha_2 p_2(x_n) + \cdots + \alpha_m p_m(x_n) + \epsilon_n \end{aligned}$$

which, when written in matrix form, becomes:

$$\mathbf{y} = \mathbf{P}\boldsymbol{\alpha} + \boldsymbol{\epsilon} \quad (16.204)$$

The matrix  $\mathbf{P}$  consists of vectors of the orthogonal polynomials computed at the discrete values  $x_k$ , just as we showed in Example 16.10 for the Gram polynomials. The least-squares solution to this equation,

$$\hat{\boldsymbol{\alpha}} = (\mathbf{P}^T \mathbf{P})^{-1} \mathbf{P}^T \mathbf{y}, \quad (16.205)$$

looks very much like what we have seen before, until we recall that, as a result of the orthogonality of the constituent elements of  $\mathbf{P}$ , the matrix  $\mathbf{P}^T \mathbf{P}$  is diagonal, with elements  $\psi_i^2$ , because all the off-diagonal terms vanish identically (see Eq (16.197) and Example 16.10)). As a result, the expression in Eq (16.205) is nothing but a collection of  $n$  isolated algebraic equations,

$$\hat{\alpha}_i = \frac{\sum_{k=1}^n p_i(x_k) y(x_k)}{\psi_i^2} \quad (16.206)$$

where

$$\psi_i^2 = \sum_{k=1}^n p_i(x_k) p_i(x_k); i = 0, 1, \dots, m. \quad (16.207)$$

This approach has several additional advantages beyond the dramatically simplified computation:

1. Each parameter estimate,  $\hat{\alpha}_i$ , is independent of the others, and its value remains unaffected by the order chosen for the polynomial model. In other words, after obtaining the first  $m$  parameter estimates for an  $m^{th}$  order polynomial model, should we decide to increase the polynomial order to  $m + 1$ , the new parameter estimate,  $\hat{\alpha}_{m+1}$ , is simply obtained as

$$\hat{\alpha}_{m+1} = \frac{\sum_{k=1}^n p_{m+1}(x_k) y(x_k)}{\xi_{m+1}^2} \quad (16.208)$$

using the very same data set  $y(x_k)$ , and only introducing  $p_{m+1}(k)$ , a pre-computed vector of the  $(m+1)^{th}$  polynomial. All the previously obtained values for  $\hat{\alpha}_i$ ;  $i = 1, 2, \dots, m$ , remain unchanged. This is very convenient indeed. Recall that this is not the case with regular polynomial regression (see Examples 16.9 and 16.10) where a change in the order of the polynomial model mandates a change in the values estimated for the new set of parameters.

2. From earlier results, we know that the variance associated with the estimates,  $\hat{\alpha}_i$ , is given by:

$$\Sigma_{\hat{\alpha}} = (\mathbf{P}^T \mathbf{P})^{-1} \sigma^2 \quad (16.209)$$

But, by virtue of the orthogonality of the elements of the  $\mathbf{P}$  matrix, this reduces to:

$$\sigma_{\hat{\alpha}_i}^2 = \frac{\sigma^2}{\psi^2} \quad (16.210)$$

and since the value for the term  $\psi_i^2$ , defined as in Eq (16.207), is determined strictly by the placement of the “design” points,  $x_k$ , where the data is obtained, Eq (16.210) indicates that this approach makes it possible to select experimental points such as to influence the variance of the estimated parameters favorably, with obvious implications for strategic design of experiments.

3. Finally, it can be shown that  $\psi_i^2$  decreases monotonically with  $i$ , indicating that the precision with which coefficients of higher order polynomials are estimated worsens with increasing order. This is also true for regular polynomial regression, but it is not as obvious.

An example of how orthogonal polynomial regression has been used in engineering applications may be found in Kristinsson and Dumont, 1993<sup>5</sup>, and 1996<sup>6</sup>.

## 16.6 Summary and Conclusions

The fitting of simple empirical mathematical expressions to data is an activity with which many engineers and scientists are very familiar, and perhaps have been engaged in even since high school. This chapter has therefore been more or less a re-introduction of the reader to regression analysis, especially to the fundamental principles behind the mechanical computations that are now routinely carried out with computer software. We have shown regression analysis to be a direct extension of estimation to cases where the mean of the random variation in the observations is no longer constant (as in our earlier discussions) but now varies as a function of one or more independent variables. The primary problem in regression analysis is therefore

<sup>5</sup>K. Kristinsson and G. A. Dumont, “Paper Machine Cross Directional Basis Weight Control Using Gram Polynomials,” *Proceedings of the Second IEEE Conference on Control Applications*, p235 -240, September 13 - 16, 1993, Vancouver, B.C.

<sup>6</sup>K. Kristinsson and G. A. Dumont, “Cross-directional control on paper machines using Gram polynomials,” *Automatica* 32 (4) 533 - 548 (1996)

the determination of the unknown parameters contained in the functional relationship (the regression model equation), given appropriate experimental data. The primary method discussed in this chapter for carrying out this task is the method of least squares. However, when regression analysis is cast as the probabilistic estimation problem that it truly is fundamentally, one can also employ the method of maximum likelihood to determine the unknown parameters. However, this requires the explicit specification of a probability distribution for the observed random variability—something not explicitly required by the method of least squares. Still, under the normal distribution assumption, maximum likelihood estimates of the regression model parameters coincide precisely with least squares estimates (See Exercises 16.5 and 16.6).

In addition to the familiar, we have also presented some results for specialized problems, for example, when variances are not uniform across observations (weighted least squares); when the parameters are not independent but are subject to (linear) constraints (constrained least squares); when the data matrix is poorly conditioned perhaps because of collinearity (ridge regression); and when information is available sequentially (recursive least squares). Space constraints compel us to limit the illustration and application of these techniques to a handful of end-of-chapter exercises and application problems, which are highly recommended to the reader.

It bears repeating, in conclusion, that since all the computations required for regression analysis are now routinely carried out with the aid of computers, it is all the more important to concentrate on understanding the principles behind these computations, so that computer-generated results can be *interpreted* appropriately. In particular, first, the well-informed engineer should understand the implications of the following on the problem at hand:

- the results of hypothesis tests on the significance of estimated parameters;
- the  $R^2$  and  $R_{adj}^2$  values as measures of how much of the information contained in the data has been adequately explained by the regression model, and with the “expenditure” of how many significant parameters;
- the value computed for the standard error of the residuals.

These will always be computed by any regression analysis software as a matter of course. Next, other quantities such as confidence and prediction intervals, and especially *residuals*, can be generated upon request. It is highly recommended that every regression analysis be accompanied by a thorough analysis of the residuals as a matter of routine “diagnostics.” The principles—and mechanics—behind how the assumption (explicit or implicit) of the normality of residuals are validated systematically and rigorously is discussed in the next chapter as part of a broader discussion of probability model validation.

---

## REVIEW QUESTIONS

- 1.** In regression analysis, what is an independent variable and what is a dependent variable?
- 2.** In regression analysis as discussed in this chapter, which variable is deterministic and which is random?
- 3.** In regression analysis, what is a “predictor” and what is a response variable?
- 4.** Regression analysis is concerned with what tasks?
- 5.** What is the principle of least squares?
- 6.** In simple linear regression, what is a one-parameter model; what is a two-parameter model?
- 7.** What are the two main assumptions underlying regression analysis?
- 8.** In simple linear regression, what are the “normal equations” and how do they arise?
- 9.** In simple linear regression, under what conditions are the least squares estimates identical to the maximum likelihood estimates?
- 10.** In regression analysis, what are residuals?
- 11.** What does it mean that OLS estimators are unbiased?
- 12.** Why is the confidence interval around the regression line curved? Where is the interval narrowest?
- 13.** What does hypothesis testing entail in linear regression? What is  $H_0$  and what is  $H_a$  in this case?
- 14.** What is the difference between using the regression line to estimate mean responses and using it to predict a new response?
- 15.** Why are prediction intervals consistently wider than confidence intervals?
- 16.** What is  $R^2$  and what is its role in regression?
- 17.** What is  $R_{adj}^2$  and what differentiates it from  $R^2$ ?
- 18.** Is an  $R^2$  value close to 1 always indicative of a good regression model? By the same token, is an  $R_{adj}^2$  value close to 1 always indicative of a good regression model?

- 19.** In the context of simple linear regression, what is an  $F$ -test used for?
- 20.** What is the connection between  $R^2$ , the coefficient of determination, and the correlation coefficient?
- 21.** If a regression model represents a data set adequately, what should we expect of the residuals?
- 22.** What does residual analysis allow us to do?
- 23.** What activities are involved in residual analysis?
- 24.** What are standardized residuals?
- 25.** Why is it recommended for residual plots to be based on standardized residuals?
- 26.** The term “linear” in linear regression refers to what?
- 27.** As far as regression is concerned, how does one determine whether the problem is linear or nonlinear?
- 28.** What is an “intrinsically linear” model?
- 29.** In employing variable transformations to convert nonlinear regression problems to linear ones, what important issue should be taken into consideration?
- 30.** What is multiple linear regression?
- 31.** What is the “hat” matrix and what is its role in multiple linear regression?
- 32.** What are some reasons for using weights in regression problems?
- 33.** What is constrained least squares and what class of problems require this approach?
- 34.** What is ridge regression and under what condition is it recommended?
- 35.** What is the principle behind recursive least squares?
- 36.** What is polynomial regression?
- 37.** What is special about orthogonal polynomial regression?
- 38.** What is the orthogonal basis function expansion result and what are its implications for polynomial regression?

## EXERCISES

**16.1** Given the one-parameter model,

$$y_i = x_i\theta + \epsilon_i$$

where  $\{y_i\}_{i=1}^n$  is the specific sample data set, and  $\epsilon_i$ , the random error component, has zero mean and variance  $\sigma^2$ , it was shown in Eq (19.49) that the least squares estimate of the parameter  $\theta$  is

$$\hat{\theta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$$

- (i) Show that this estimate is unbiased for  $\theta$ , i.e.,  $E(\hat{\theta}) = \theta$
- (ii) Determine  $Var(\hat{\theta})$ .

**16.2** Consider the random sample,  $Y_1, Y_2, \dots, Y_n$  drawn from a population characterized by a single, constant parameter,  $\theta$ , the population mean, so that the random variable  $Y$  may then be written as:

$$Y_i = \theta + \epsilon_i$$

Determine  $\hat{\theta}_w$ , the weighted least squares estimate of  $\theta$  by solving the minimization problem

$$\min_{\theta} S_w(\theta) = \sum_{i=1}^n W_i(Y_i - \theta)^2$$

and hence establish the results in Eqs (16.11) and (16.12).

**16.3** For the one-parameter model,

$$y_i = x_i\theta + \epsilon_i$$

where  $\{y_i\}_{i=1}^n$  is the specific sample data set, and  $\epsilon_i$ , the random error component, has zero mean and variance  $\sigma^2$ ,

- (i) Determine  $\hat{\theta}_w$ , the weighted least squares estimate of  $\theta$  by solving the minimization problem

$$\min_{\theta} S_w(\theta) = \sum_{i=1}^n W_i[y_i - (x_i\theta)]^2$$

and compare it to the ordinary least squares estimate obtained in Eq (19.49).

- (ii) Show that  $E(\hat{\theta}_w) = \theta$ .
- (iii) Determine  $Var(\hat{\theta}_w)$ .

**16.4** For the two-parameter model,

$$y_i = \theta_0 + \theta_1 x_i + \epsilon_i$$

where, as in Exercise 16.3,  $\{y_i\}_{i=1}^n$  is the sample data, and  $\epsilon_i$ , the random error, has zero mean and variance  $\sigma^2$ ,

(i) Determine  $\boldsymbol{\theta}_\omega = (\hat{\theta}_{0\omega}, \hat{\theta}_{1\omega})$ , the weighted least squares estimate of  $\boldsymbol{\theta} = (\theta_0, \theta_1)$ , by solving the minimization problem

$$\min_{\theta_0, \theta_1} S_w(\hat{\theta}_0, \hat{\theta}_1) = \sum_{i=1}^n W_i [y_i - (\theta_0 + \theta_1 x_i)]^2$$

and compare these to the ordinary least squares estimates obtained in Eqs (16.38) and (16.37).

- (ii) Show that  $E(\hat{\theta}_{0\omega}) = \theta_0$ ; and  $E(\hat{\theta}_{1\omega}) = \theta_1$ .
- (iii) Determine  $Var(\hat{\theta}_{0\omega})$  and  $Var(\hat{\theta}_{1\omega})$ .

**16.5** Let  $Y_1, Y_2, \dots, Y_n$  be a random sample from a Gaussian distribution with mean  $\eta$  and variance,  $\sigma^2$ ; i.e.,  $Y \sim N(\eta(x, \boldsymbol{\theta}), \sigma^2)$ .

- (i) For the one-parameter model where

$$\eta = \theta x$$

determine the maximum likelihood estimate of the parameter  $\theta$  and compare it to the least squares estimate in Eq (19.49).

- (ii) When the model is

$$\eta = \theta_0 + \theta_1 x$$

(the two-parameter model), determine the maximum likelihood estimates of the parameters  $\theta_0$  and  $\theta_1$ ; compare these to the corresponding least squares estimates in Eqs (16.38) and (16.37).

**16.6** Let each individual observation,  $Y_i$  be an independent sample from a Gaussian distribution with mean  $\eta$  and variance,  $\sigma_i^2$ , i.e.,  $Y_i \sim N(\eta(x, \boldsymbol{\theta}), \sigma_i^2)$ ;  $i = 1, 2, \dots, n$ ; where the variances are not necessarily equal.

- (i) Determine the maximum likelihood estimate of the parameter  $\theta$  in the one-parameter model,

$$\eta = \theta x$$

and show that it has the form of a *weighted* least squares estimate. What are the weights?

- (ii) Determine the maximum likelihood estimates of the parameters  $\theta_0$  and  $\theta_1$  in the two-parameter model,

$$\eta = \theta_0 + \theta_1 x$$

Show that these are also similar to weighted least squares estimates. What are the weights in this case?

**16.7** From the definitions of the estimators for the parameters in the two-parameter model given in Eqs (16.46) and (16.47), i.e.,

$$\begin{aligned}\Theta_1 &= \frac{S_{xy}}{S_{xx}} \\ \Theta_0 &= \bar{Y} - \Theta_1 \bar{x}\end{aligned}$$

obtain expressions for the respective variances for each estimator and hence establish the results given in Eqs (16.55) and (16.56).

**16.8** A fairly common mistake in simple linear regression is the use of the one-parameter model (where the intercept is implicitly set to zero) in place of the more general two-parameter model, thereby losing the flexibility of estimating an intercept which may or may not be zero. When the true intercept in the data is non-zero, such a mistake will lead to an error in the estimated value of the single parameter, the slope  $\theta$ , because the least squares criterion has no option but to honor the implicit constraint forcing the intercept to be zero. This will compromise the estimate of the true slope. The resulting estimation error may be quantified explicitly as follows.

First, show that the relationship between  $\hat{\theta}_1$ , the estimated slope in the two-parameter model, and  $\hat{\theta}$ , the estimated slope in the one-parameter model, is:

$$\hat{\theta}_1 = \left( \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \right) \hat{\theta} - \left( \frac{n\bar{x}^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \right) \frac{\bar{y}}{\bar{x}}$$

so that the two slopes are the same if and only if

$$\hat{\theta} = \frac{\bar{y}}{\bar{x}} \quad (16.211)$$

which will be the case when the intercept is truly zero. Next, show from here that

$$\hat{\theta} = \alpha \left( \frac{\bar{y}}{\bar{x}} \right) + (1 - \alpha) \hat{\theta}_1 \quad (16.212)$$

with  $\alpha = n\bar{x}^2 / \sum_{i=1}^n x_i^2$ , indicating clearly the least squares compromise—a weighted average of  $\hat{\theta}_1$ , (the true slope when the intercept *is not zero*), and  $\bar{x}/\bar{y}$  (the true slope when the intercept *is actually zero*).

Finally, show that the estimation error,  $e_\theta = \hat{\theta}_1 - \hat{\theta}$  will be given by:

$$e_\theta = \hat{\theta}_1 - \hat{\theta} = \alpha \left( \hat{\theta}_1 - \frac{\bar{y}}{\bar{x}} \right) \quad (16.213)$$

**16.9** By defining as  $S_{yy}$  the total variability present in the data, i.e.,  $\sum_{i=1}^n (y_i - \bar{y})^2$  (see Eq (16.32)), and by rearranging this as follows:

$$S_{yy} = \sum_{i=1}^n [(y_i - \hat{y}_i) - (\bar{y} - \hat{y}_i)]^2$$

expand and simplify this expression to establish the important result in Eq (16.89),

$$\begin{aligned} \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ \text{or } S_{yy} &= SS_R + SS_E \end{aligned}$$

**16.10** Identify which of the following models presents a linear or nonlinear regression problem in estimating the unknown parameters,  $\theta_i$ .

- (i)  $Y = \theta_0 + \theta_2 x^{\theta_3} + \epsilon$
- (ii)  $Y = \theta_0 + \frac{\theta_1}{x} + \epsilon$
- (iii)  $Y = \theta_0 + \theta_1 e^x + \theta_2 \sin x + \theta_3 \sqrt{x} + \epsilon$
- (iv)  $Y = \theta_0 e^{-\theta_2 x} + \epsilon$
- (v)  $Y = \theta_0 x_1^{\theta_1} x_2^{\theta_2} + \epsilon$

**16.11** The following models, sampled from various branches of science and engineering, are nonlinear in the unknown parameters. Convert each into a linear regression model; indicate an explicit relationship between the original parameters and the transformed ones.

(i) **Antoine's Equation:** Vapor pressure,  $P^{vap}$ , as a function of temperature,  $T$ :

$$P^{vap} = e^{\theta_0 - \frac{\theta_1}{T + \theta_3}}$$

(ii) **Cellular growth rate** (exponential phase):  $N$ , number of cells in the culture, as a function of time,  $t$ .

$$N = \theta_0 e^{\theta_1 t}$$

(iii) **Kleiber's law of bioenergetics:** Resting energy expenditure,  $Q_0$ , as a function of an animal's mass,  $M$ :

$$Q_0 = \theta_0 M^{\theta_1}$$

(iv) **Gilliland-Sherwood correlation:** Mass-transfer in falling liquid films in terms of  $Sh$ , Sherwood number, as a function of two other dimensionless numbers,  $Re$ , the Reynolds number, and  $Sc$ , the Schmidt number:

$$Sh = \theta_0 Re^{\theta_1} Sc^{\theta_2}$$

**16.12** Establish that the “hat” matrix,

$$\mathbf{H} = \mathbf{X} \left( \mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T$$

is idempotent. Establish that  $(\mathbf{I} - \mathbf{H})$  is also idempotent. As a result, establish that not only is  $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ , but

$$\mathbf{H}\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$$

Similarly, not only is  $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y}$ , but

$$(\mathbf{I} - \mathbf{H})\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

where  $\mathbf{e}$  is the residual vector defined in Eq (16.160).

**16.13** The angles between three celestial bodies that aligned to form a triangle in a plane at a particular point in time have been measured as  $y_1 = 91^\circ$ ,  $y_2 = 58^\circ$  and  $y_3 = 33^\circ$ . Since the measurement device cannot determine these angles without error, the results do not add up to  $180^\circ$  as they should; but arbitrarily forcing the numbers to add up to  $180^\circ$  is *ad-hoc* and undesirable. Formulate the problem instead as a constrained least squares problem

$$y_i = \theta_i + \epsilon_i; \quad i = 1, 2, 3,$$

subject to the constraint:

$$\theta_1 + \theta_2 + \theta_3 = 180$$

and determine the least squares estimate of these angles. Confirm that these estimates add up to  $180^\circ$ .

**16.14** When the data matrix  $\mathbf{X}$  in the multiple regression equation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$$

is poorly conditioned, it was recommended in the text that the ridge regression estimate:

$$\hat{\boldsymbol{\theta}}_{RR} = (\mathbf{X}^T \mathbf{X} + k^2 \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$$

might provide a more reliable estimate.

- (i) Given that  $E(\epsilon) = 0$ , show that  $\hat{\boldsymbol{\theta}}_{RR}$  is *biased* for the parameter vector  $\boldsymbol{\theta}$ .
- (ii) Given  $E(\epsilon\epsilon^T) = \boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$  as the error covariance matrix, determine the covariance matrix for the ridge regression estimate,

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_{RR}} = \text{Var}(\hat{\boldsymbol{\theta}}_{RR}) = E[(\hat{\boldsymbol{\theta}}_{RR} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_{RR} - \boldsymbol{\theta})^T]$$

and compare it to the covariance matrix for the regular least squares estimate shown in Eq (16.151).

**16.15** An experimental investigation over a narrow range of the independent variable  $x$  yielded the following paired data:  $(x_1 = 1.9, y_1 = 4.89)$ ;  $(x_2 = 2.0, y_1 = 4.95)$  and  $(x_3 = 2.1, y_3 = 5.15)$ . It is postulated that the relationship between  $y$  and  $x$  is the two-parameter model

$$y = \theta_0 + \theta_1 x + \epsilon$$

- (i) First, confirm that the true values of the unknown parameters are  $\theta_0 = 1$  and  $\theta_1 = 2$ , by computing for each indicated value of  $x_i$ , the theoretical responses  $\eta_i$  given by:

$$\eta_i = 1 + 2x_i$$

and confirming that within the limits of experimental error, the predicted  $\eta_i$  matches the corresponding experimental response,  $y_i$ .

- (ii) Next, obtain the data matrix  $\mathbf{X}$  and confirm that  $\mathbf{X}^T \mathbf{X}$  is close to being singular, by computing the determinant of this matrix.
- (iii) Determine the least squares estimates of  $\theta_0$  and  $\theta_1$ ; compare these to the true values given in (i).
- (iv) Next compute three different sets of ridge regression estimates for these parameters using the the values  $k^2 = 2.0, 1.0$ , and  $0.5$ . Compare these ridge regression estimates to the true value.
- (v) Plot on the same graph, the data, the regular least squares fit and the best ridge regression fit.

**16.16** Consider the data in the table below.

| $x$   | $y$     |
|-------|---------|
| -1.00 | -1.9029 |
| -0.75 | -0.2984 |
| -0.50 | 0.4047  |
| -0.25 | 0.5572  |
| 0.00  | 0.9662  |
| 0.25  | 2.0312  |
| 0.50  | 3.2286  |
| 0.75  | 5.7220  |
| 1.00  | 10.0952 |

Fit a series of polynomial models of increasing complexity as follows. First fit the linear, two-parameter model,

$$y = \theta_0 + \theta_1 x + \epsilon$$

next fit the quadratic model,

$$y = \theta_0 + \theta_1 x + \theta_2 x^2 + \epsilon$$

and then fit the cubic model,

$$y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \epsilon$$

and finally fit the quartic model,

$$y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4 + \epsilon$$

In each case, note the values of the estimates for each parameter, check the  $R^2$  and  $R_{adj}^2$  values and the significance of each model parameter. Plot each model fit to the data and select the most appropriate model.

**16.17** Refer to the data table in Exercise 16.16. This time use Gram orthogonal polynomials with  $n = 7$  to obtain linear, quadratic, cubic, and quartic fits to this data sequentially as in Exercise 16.16. Compare the coefficients of the orthogonal polynomial fits among themselves and then compare them to the coefficients obtained in Exercise 16.16.

## APPLICATION PROBLEMS

**16.18** A predictive model is sometimes evaluated by plotting its predictions directly against the corresponding experimental data in an  $(x, y)$  plot: if the model predicts the data adequately, a regression line fit should be a 45 degree line with slope 1 and intercept 0; the residuals should appear as a random sequence of numbers that are independent, normally distributed, and with zero mean and variance close to measurement error variance. This technique is to be used to evaluate two models of multicomponent transport as follows.

In Kerkhof and Geboers, (2005)<sup>7</sup>, the authors presented a new approach to modeling multicomponent transport that is purported to yield more accurate predictions than previously available models. The table below shows experimentally determined viscosity ( $10^{-5} \text{ Pa.s}$ ) of 12 different gas mixtures and the corresponding values predicted by two models: (i) the classical Hirschfelder-Curtiss-Bird (HCB) model<sup>8</sup>, and (ii) their new (KG) model.

<sup>7</sup>Kerkhof, P.J.A.M, and M.A.M. Geboers, (2005). "Toward a unified theory of isotropic molecular transport phenomena," *AIChE Journal*, 51, (1), 79–121

<sup>8</sup>Hirschfelder J.O., C.F. Curtiss, and R.B. Bird (1964). *Molecular Theory of Gases and Liquids*. 2nd printing. J. Wiley, New York, NY.

| Viscosity, ( $10^{-5} \text{ Pa.s}$ ) |                 |                |
|---------------------------------------|-----------------|----------------|
| Experimental Data                     | HCB Predictions | KG Predictions |
| 2.740                                 | 2.718           | 2.736          |
| 2.569                                 | 2.562           | 2.575          |
| 2.411                                 | 2.429           | 2.432          |
| 2.504                                 | 2.500           | 2.512          |
| 3.237                                 | 3.205           | 3.233          |
| 3.044                                 | 3.025           | 3.050          |
| 2.886                                 | 2.895           | 2.910          |
| 2.957                                 | 2.938           | 2.965          |
| 3.790                                 | 3.752           | 3.792          |
| 3.574                                 | 3.551           | 3.582          |
| 3.415                                 | 3.425           | 3.439          |
| 3.470                                 | 3.449           | 3.476          |

- (i) Treating the KG model prediction as the independent variable and the experimental data as the response variable, fit a two-parameter model and *thoroughly* evaluate the regression results, the parameter estimates, their significance,  $R^2$  and  $R_{adj}^2$  values, and the residuals. Plot the regression line along with a 95% confidence interval around the regression line. In light of this regression analysis, provide your opinion about the author's claim that their model provides an "excellent agreement" with the data.
- (ii) Repeat (i) for the HCB model. In light of your results here and in (i), comment on whether or not the KG model can truly be said to provide better predictions than the HCB model.

**16.19** In an attempt to quantify a possible relationship between the amount of fire damage caused by residential fires and the distance from the residence to the closest fire station, the following data were acquired from a random sample of 12 recent fires.

| Distance from Fire Station<br>$x$ (miles) |                                 | Distance from Fire Station<br>$x$ (miles) |                                 |
|---|---------------------------------|---|---------------------------------|
|   | Fire damage<br>$y$ (\$ $10^3$ ) |   | Fire damage<br>$y$ (\$ $10^3$ ) |
| 1.8                                       | 17.8                            | 5.5                                       | 36.0                            |
| 4.6                                       | 31.3                            | 3.0                                       | 22.3                            |
| 0.7                                       | 14.1                            | 4.3                                       | 31.3                            |
| 3.4                                       | 26.2                            | 1.1                                       | 17.3                            |
| 2.3                                       | 23.1                            | 3.1                                       | 27.5                            |
| 2.6                                       | 19.6                            | 2.1                                       | 24.0                            |

- (i) Postulate an appropriate model, estimate the model parameters and evaluate the model fit.
- (ii) An insurance company wishes to use this model to estimate the expected fire damage to two new houses, house  $A$  that is being built at a distance of 5 miles from the nearest fire station, and house  $B$ , 3 miles from the same fire station. Determine these estimates along with appropriate uncertainty intervals.
- (iii) Is it "safe" to use this model to predict the fire damage to a house  $C$  that is being built 6 miles from the nearest fire station? Regardless of your answer, provide a prediction and an appropriate uncertainty interval.

**16.20** Refer to Problem 16.19. Now consider that three new residential fires have

occurred and the following additional data set has become available at the same time.

| Distance from<br>Fire Station |  | Fire damage  |
|-------------------------------|--|--------------|
| $x$ (miles)                   |  | $y$ (\$10^3) |
| 3.8                           |  | 26.1         |
| 4.8                           |  | 36.4         |
| 6.1                           |  | 43.2         |

- (i) Use recursive least squares to adjust the previously obtained set of parameter estimates in light of this new information. By how much have the parameters changed?
- (ii) Recalculate the estimated expected fire damage values to houses  $A$  and  $B$  in light of the new data; compare these values to the corresponding values obtained in Exercise 16.19. Have these values changed by amounts that may be considered practically important?
- (iii) With this new data, is it “safe” to use the updated model to predict the fire damage to house  $C$ ? Predict this fire damage amount.

**16.21** In Ogunnaike, (2006)<sup>9</sup>, the data in the following table was used to characterize the extent of DNA damage,  $\lambda$ , experienced by cells exposed to radiation of dose  $\gamma$  (Gy), with the power-law relationship,

$$\lambda = \theta_0 \gamma^{\theta_1}$$

| Radiation Dose | Extent of<br>DNA damage, $\lambda$ |
|----------------|------------------------------------|
| $\gamma$ (Gy)  |                                    |
| 0.00           | 0.05                               |
| 0.30           | 1.30                               |
| 2.50           | 2.10                               |
| 10.0           | 3.10                               |

- (i) Using an appropriate variable transformation, estimate the unknown model parameters. Are both model parameters significant at the 95% level? What do the  $R^2$  and  $R_{adj}^2$  values suggest about the model fit to the data?
- (ii) The extent of DNA damage parameter,  $\lambda$ , is in fact the Poisson random variable parameter,  $\lambda$ ; in this case, it represents what amounts to the mean number of strand breaks experienced by a cell in a population of cells exposed to gamma radiation. If a cell experiences a total of  $2n$  strand breaks (or  $n$  double-strand breaks),  $n$  pulses of the DNA-repair protein  $p53$  will be observed in response. Use the model obtained in (i) to determine the probability that upon exposure to radiation of 5 Gy, a cell undergoes DNA damage that will cause the cell’s DNA-damage repair system to respond with 3 or more pulses of  $p53$ .

**16.22** Consider a manufacturing process for making pistons from metal ingots in which each ingot produces enough material for 1,000 pistons. Occasionally, a piston cracks while cooling after being forged. Previous research has indicated that, in a batch of 1,000 pistons, the average number of pistons that develop a crack during

<sup>9</sup>Ogunnaike, B. A. (2006). “Elucidating the digital control mechanism for DNA damage repair with the p53-Mdm2 System: Single cell data analysis and ensemble modeling,” *J. Roy. Soc. Interface*. 3, 175-184.

cooling is dependent on the purity of the ingots. Ingots of known purity were forged into pistons, and the average number of cracked pistons per batch was recorded in the table below.

| Purity, $x$                   | 0.94 | 0.95 | 0.96 | 0.97 | 0.98 | 0.99 |
|-------------------------------|------|------|------|------|------|------|
| Ave # of Cracked Pistons, $y$ | 4.8  | 4.6  | 3.9  | 3.3  | 2.7  | 2.0  |

Over the small range of purity, the dependence of  $y$ , the average number of cracked pistons per batch, on purity  $x$ , may assumed to be linear, i.e.,

$$y = \theta_1 x + \theta_0 \quad (16.214)$$

- (i) Estimate the parameters  $\theta_0$  and  $\theta_1$ .
- (ii) A steel mill claims to have produced 100 ingots with a purity of 96% (i.e.,  $x = 0.96$ ), 5 of which are used to manufacture 1,000 pistons. The number of cracked pistons from each ingot/batch is recorded in the table below.

| Ingot #              | 1 | 2 | 3 | 4 | 5 |
|----------------------|---|---|---|---|---|
| # of Cracked Pistons | 4 | 6 | 3 | 6 | 6 |

Knowing the average number of cracked pistons per batch expected for ingots of 96% purity via Eq (16.214) and the parameters found in part (i), determine if the steel mill's purity claim is reasonable, based on the sample of 5 ingots.

- (iii) Repeat part (ii) assuming that 20 ingots from the mill were tested instead of 5 but that the mean and variance for the 20 ingots are the same as was calculated for the sample of 5 ingots in part (ii).

**16.23** The table below, first introduced in Chapter 12, shows city and highway gasoline mileage ratings, in miles per gallon (mpg), for 20 types of two-seater automobiles, complete with engine characteristics, capacity (in liters) and number of cylinders.

- (i) Obtain an appropriate regression models relating the number of cylinders (as  $x$ ) to highway gas mileage and to city gas mileage (as  $y$ ). At the 95% significance level, is there a difference between the parameters of these different models?
- (ii) By analyzing the residuals, do these models provide reasonable explanations of how the number of cylinders a car engine has affects the gas mileage, either in the city or on the highway?

|    | Car Type and Model      | Eng Capacity (Liters) | # Cylinders | City mpg | Highway mpg |
|----|-------------------------|-----------------------|-------------|----------|-------------|
| 1  | Aston Marton V8 Vantage | 4.3                   | 8           | 13       | 20          |
| 2  | Audi R8                 | 4.2                   | 8           | 13       | 19          |
| 3  | Audi TT Roadster        | 2.0                   | 4           | 22       | 29          |
| 4  | BMW Z4 3.0i             | 3.0                   | 6           | 19       | 28          |
| 5  | BMW Z4 Roadster         | 3.2                   | 6           | 15       | 23          |
| 6  | Bugatti Veyron          | 8.0                   | 16          | 8        | 14          |
| 7  | Caddilac XLR            | 4.4                   | 8           | 14       | 21          |
| 8  | Chevrolet Corvette      | 7.0                   | 8           | 15       | 24          |
| 9  | Dodge Viper             | 8.4                   | 10          | 13       | 22          |
| 10 | Ferrari 599 GTB         | 5.9                   | 12          | 11       | 15          |
| 11 | Honda S2000             | 2.2                   | 4           | 18       | 25          |
| 12 | Lamborghini Murcielago  | 6.5                   | 12          | 8        | 13          |
| 13 | Lotus Elise/Exige       | 1.8                   | 4           | 21       | 27          |
| 14 | Mazda MX5               | 2.0                   | 4           | 21       | 28          |
| 15 | Mercedes Benz SL65 AMG  | 6.0                   | 12          | 11       | 18          |
| 16 | Nissan 350Z Roadster    | 3.5                   | 6           | 17       | 24          |
| 17 | Pontiac Solstice        | 2.4                   | 4           | 19       | 25          |
| 18 | Porsche Boxster-S       | 3.4                   | 6           | 18       | 25          |
| 19 | Porsche Cayman          | 2.7                   | 6           | 19       | 28          |
| 20 | Saturn SKY              | 2.0                   | 4           | 19       | 28          |

**16.24** Refer to the data set in Problem 16.23. Repeat the analysis this time for engine capacity. Examine the residuals and comment on any unusual observations.

**16.25** The data in the table below shows the size of a random sample of 10 homes (in square feet) located in the Mid-Atlantic region of the US, and the corresponding amount of electricity used (KW-hr) monthly in each home.

From a scatter plot of the data, postulate an appropriate regression model and estimate the parameters. Comment on the significance of the estimated parameters. Investigate the residuals and comment on the model fit to the data. What do the model parameters signify about how the size of a home in this region of the US influences the amount of electricity used?

| Home Size<br><i>x</i> (sq. ft) | Electricity<br>Usage <i>y</i> (KW-hr) |
|--------------------------------|---------------------------------------|
| 1290                           | 1182                                  |
| 1350                           | 1172                                  |
| 1470                           | 1264                                  |
| 1600                           | 1493                                  |
| 1710                           | 1571                                  |
| 1840                           | 1711                                  |
| 1980                           | 1804                                  |
| 2230                           | 1840                                  |
| 2400                           | 1956                                  |
| 2930                           | 1954                                  |

**16.26** Some of the earliest contributions to chemical engineering science occurred in the form of correlations that shed light on mass and heat transfer in liquids and gases. These correlations were usually presented in the form of dimensionless numbers, combinations of physical variables that enable general characterization of transport and other phenomena in a manner that is independent of specific physical dimensions

of the equipment in question. One such correlation regarding mass-transfer in falling liquid films is due to Gilliland and Sherwood, (1934)<sup>10</sup>. It relates the Sherwood number,  $Sh$ , (a dimension number representing the ratio of convective to diffusive mass transport) to two other dimensionless numbers: the Reynolds number,  $Re$ , (a dimensionless number that gives a measure of the ratio of inertial forces to viscous forces) and the Schmidt number,  $Sc$  (the ratio of momentum diffusivity (viscosity) to mass diffusivity; it represents the relative ease of molecular momentum and mass transfer).

A sample of data from the original large set (of almost 400 data points!) is shown in the table below. From a postulated model,

$$Sh = \theta_0 Re^{\theta_1} Sc^{\theta_2}$$

- (i) Obtain estimates of the parameters  $\theta_0, \theta_1$  and  $\theta_2$  and compare them to the result in the original publication, respectively, 0.023; 0.830; 0.440.
- (ii) Using parameters estimated in (i), plot the data in terms of  $\log(Sh/Sc^{\hat{\theta}_2})$  vs  $\log Re$  along with your regression model and comment on the observed fit.

| $Sh$ | $Re$  | $Sc$  |
|------|-------|-------|
| 43.7 | 10800 | 0.600 |
| 21.5 | 5290  | 0.600 |
| 42.9 | 7700  | 1.610 |
| 19.8 | 2330  | 1.610 |
| 24.2 | 3120  | 1.800 |
| 88.0 | 14400 | 1.800 |
| 93.0 | 16300 | 1.830 |
| 70.6 | 13000 | 1.830 |
| 32.3 | 4250  | 1.860 |
| 56.0 | 8570  | 1.860 |
| 51.6 | 6620  | 1.875 |
| 50.7 | 8700  | 1.875 |
| 26.1 | 2900  | 2.160 |
| 41.3 | 4950  | 2.160 |
| 92.8 | 14800 | 2.170 |
| 54.2 | 7480  | 2.170 |
| 65.5 | 9170  | 2.260 |
| 38.2 | 4720  | 2.260 |

**16.27** For efficient and profitable operation, (especially during the summer months), electrical power companies need to predict, as precisely as possible, “Peak power load” ( $P^*$ ), defined as *daily maximum amount of power required to meet demand*. The inability to predict  $P^*$  accurately and to provide sufficient power to meet the indicated demand is responsible in part for many blackouts/brownouts.

The data shown in the table below is a random sample of 30 daily high temperatures ( $T^{\circ}\text{F}$ ) and corresponding  $P^*$  (in megawatts) acquired between the months of May and August in a medium-sized city.

---

<sup>10</sup>Gilliland, E.R. and Sherwood, T.K., 1934. Diffusion of vapors into air streams. *Ind. Engng Chem.* 26, 516–523.

| Temp(°F) | $P^*$ (megawatts) | Temp(°F) | $P^*$ (megawatts) |
|----------|-------------------|----------|-------------------|
| 95       | 140.7             | 79       | 106.2             |
| 88       | 116.4             | 76       | 100.2             |
| 84       | 113.4             | 87       | 114.7             |
| 106      | 178.2             | 92       | 135.1             |
| 94       | 136.0             | 68       | 96.3              |
| 108      | 189.3             | 85       | 111.4             |
| 90       | 132.0             | 100      | 143.6             |
| 100      | 151.9             | 74       | 103.9             |
| 71       | 92.5              | 89       | 116.5             |
| 96       | 131.7             | 86       | 105.1             |
| 67       | 96.5              | 75       | 99.6              |
| 98       | 150.1             | 70       | 97.7              |
| 97       | 153.2             | 69       | 97.6              |
| 67       | 101.6             | 82       | 107.3             |
| 89       | 118.5             | 101      | 157.6             |

- (i) From the supplied data, obtain an equation expressing  $P^*$  as a function of daily high temperature ( $T$  °F) and comment on your fit.  
(ii) Use the equation obtained in (i) to predict  $P^*$  for three different temperatures:  $T = 65$ ;  $T = 85$ ;  $T = 110$ . Is it “safe” to use the model equation for these predictions? Why, or why not?  
(iii) Predict the range of  $P^*$  corresponding to the following 2°F ranges in temperature:

$$68 \leq T_1 \leq 70; 83 \leq T_2 \leq 85; 102 \leq T_3 \leq 104$$

- (iv) If it is desired to predict  $P^*$  to within  $\pm 2\%$  of nominal value, find the maximum range of uncertainty in daily high temperature forecasts that can be tolerated for each of the following three cases:  $T_l = 69$ ;  $T_m = 84$ ;  $T_h = 103$ .

**16.28** The data in the table below, from Fern, (1983)<sup>11</sup> is to be used to calibrate a near-infra red instrument so that its reflectance measurements (at a specified wavelength) can be used to infer the protein content in wheat.

| Protein Content % | Observed Reflectance | Protein Content % | Observed Reflectance |
|-------------------|----------------------|-------------------|----------------------|
| $y$               | $x$                  | $y$               | $x$                  |
| 9.23              | 386                  | 10.57             | 443                  |
| 8.01              | 383                  | 10.23             | 450                  |
| 10.95             | 353                  | 11.87             | 467                  |
| 11.67             | 340                  | 8.09              | 451                  |
| 10.41             | 371                  | 12.55             | 524                  |
| 9.51              | 433                  | 8.38              | 407                  |
| 8.67              | 377                  | 9.64              | 374                  |
| 7.75              | 353                  | 11.35             | 391                  |
| 8.05              | 377                  | 9.70              | 353                  |
| 11.39             | 398                  | 10.75             | 445                  |
| 9.95              | 378                  | 10.75             | 383                  |
| 8.25              | 365                  | 11.47             | 404                  |

<sup>11</sup>Fern, A.T., (1983). “A misuse of ridge regression in the calibration of a near-infra red reflectance instrument,” *Applied Statistics*, 32, 73–79

Obtain an expression for the calibration line relating the protein content to the reflectance measurement. Determine the significance of the parameters. Plot the data, model line and the 95% confidence and prediction intervals. Comment objectively on how useful you expect the calibration line to be.

**16.29** The following data set, obtained in an undergraduate fluid mechanics lab experiment, shows actual air flow rate measurements, determined at room temperature, 25 °C, and 1 atmosphere of pressure, along with corresponding rotameter readings.

| Rotameter<br>Reading | Air Flow<br>rate cc/sec |
|----------------------|-------------------------|
| $x$                  | $y$                     |
| 20                   | 15.5                    |
| 40                   | 38.3                    |
| 60                   | 50.2                    |
| 80                   | 72.0                    |
| 100                  | 111.1                   |
| 120                  | 115.4                   |
| 140                  | 139.0                   |

- (i) Determine an appropriate equation that can be used to calibrate the rotameter and from which actual air flow rates can be determined for any given rotameter reading. From the significance of the parameters, the  $R^2$  and  $R_{adj}^2$  values, is this a reliable expression to use as a calibration equations? From an appropriate analysis of the residuals, comment on how carefully the experimental data were determined.
- (ii) Plot the data and the regression equation along with 95% confidence interval and prediction interval bands.
- (iii) Determine, along with 95% confidence intervals, the expected value of the air flow rates for rotameter readings of 70, 75, 85, 90, and 95.

**16.30** The data shown in the table below, from Beck and Arnold, (1977)<sup>12</sup>, shows five samples of thermal conductivity of a steel alloy as a function of temperature. The standard deviation,  $\sigma_i$ , associated with each measurement varies as indicated.

| Sample<br>$i$ | Temperature<br>$x_i$ (°C) | Thermal<br>Conductivity<br>$k_i$ W/m·°C | Standard<br>Deviation<br>$\sigma_i$ |
|---------------|---------------------------|---|-------------------------------------|
| 1             | 100                       | 36.3                                    | 0.2                                 |
| 2             | 200                       | 36.3                                    | 0.3                                 |
| 3             | 300                       | 34.6                                    | 0.5                                 |
| 4             | 400                       | 32.9                                    | 0.7                                 |
| 5             | 600                       | 31.2                                    | 1.0                                 |

Over the indicated temperature range, the thermal conductivity varies linearly with temperature; therefore the two-parameter model is deemed appropriate, i.e.,

$$k_i = \theta_0 + \theta_1 x_i + \epsilon_i$$

---

<sup>12</sup>J.V. Beck and K. J. Arnold, (1977). *Parameter Estimation in Engineering and Science*, J. Wiley, NY, p209.

- (i) Determine the weighted least squares estimates of the parameters  $\theta_0$  and  $\theta_1$  using as weights,  $w_i = 1/\sigma_i$ .
- (ii) Determine ordinary least squares estimates of the same parameters using no weights. Compare these estimates.
- (iii) Plot the data along with the two regression equations obtained above. Which one fits the data better?

**16.31** The data table below is typical of standard tables of thermophysical properties of liquids and gases used widely in chemical engineering practice, especially in process simulation. This specific data set shows the temperature dependence of the heat capacity  $C_p$  of methylcyclohexane.

| Temperature Kelvin | Heat Capacity $C_p, KJ/kg - K$ | Temperature Kelvin | Heat Capacity $C_p, KJ/kg - K$ |
|--------------------|--------------------------------|--------------------|--------------------------------|
| 150                | 1.426                          | 230                | 1.627                          |
| 160                | 1.447                          | 240                | 1.661                          |
| 170                | 1.469                          | 250                | 1.696                          |
| 180                | 1.492                          | 260                | 1.732                          |
| 190                | 1.516                          | 270                | 1.770                          |
| 200                | 1.541                          | 280                | 1.801                          |
| 210                | 1.567                          | 290                | 1.848                          |
| 220                | 1.596                          | 300                | 1.888                          |

First fit a linear model

$$C_p = \theta_0 + \theta_1 T$$

check the significance of the parameters, the residuals, and the  $R^2$  and  $R_{adj}^2$  values. Then fit the quadratic model,

$$C_p = \theta_0 + \theta_1 T + \theta_2 T^2$$

Again, check the significance of the parameters, the residuals, and the  $R^2$  and  $R_{adj}^2$  values. Finally, fit the cubic equation,

$$C_p = \theta_0 + \theta_1 T + \theta_2 T^2 + \theta_3 T^3$$

Once more, check the significance of the parameters, the residuals, and the  $R^2$  and  $R_{adj}^2$  values.

Which of the three models is more appropriate to use as an empirical relationship representing how the heat capacity of methylcyclohexane changes with temperature? Plot the data and the regression curve of the selected model, along with the 95% confidence and prediction intervals.

**16.32** The change in the bottoms temperature of a binary distillation column in response to a pulse input in the steam flow rate to the reboiler is represented by the following equation:

$$y = T - T_0 = \frac{AK}{\tau} e^{-t/\tau} \quad (16.215)$$

where  $T_0$  is the initial (steady state) temperature before the perturbation;  $A$  is the magnitude of the pulse input, idealized as a perfect delta function;  $t$  is time, and  $K$  and  $\tau$  are, respectively, the process gain, and time constant, unknown parameters

of the process when the dynamics are approximated as a “first order” system<sup>13</sup>. A “process identification” experiment performed to estimate  $K$  and  $\tau$  yielded the data in the following table, starting from an initial temperature of  $185^{\circ}C$ , and using an impulse input of magnitude  $A = 10$ .

| Time<br>$t$ (min) | Bottoms<br>$T(^{\circ}C)$ |
|-------------------|---------------------------|
| 0                 | 189.02                    |
| 1                 | 188.28                    |
| 2                 | 187.66                    |
| 3                 | 187.24                    |
| 5                 | 186.54                    |
| 10                | 185.46                    |
| 15                | 185.20                    |

Even though Eq (16.215) presents a nonlinear regression problem, it is possible, via an appropriate variable transformation, to convert it to a linear regression equation. Use an appropriate transformation and obtain an estimate of the process parameters from the provided data.

Prior to performing the experiment, an experienced plant operator stated that historically, in the operating range in question, the process parameters have been characterized as  $K \approx 2$  and  $\tau \approx 5$ . How close are these “guesstimates” to the actual estimated values?

### 16.33 Fit Antoine's equation,

$$P^{vap} = e^{\theta_0 - \frac{\theta_1}{T + \theta_3}}$$

to the data in the table below, which shows the temperature dependence of vapor pressure for Toluene.

| Toluene | $P^{vap}$ (mm Hg)<br>$T(^{\circ}C)$ |
|---------|-------------------------------------|
| 5       | -4.4                                |
| 10      | 6.4                                 |
| 20      | 18.4                                |
| 40      | 31.8                                |
| 60      | 40.3                                |
| 100     | 51.9                                |
| 200     | 69.5                                |
| 400     | 89.5                                |
| 760     | 110.6                               |
| 1520    | 136.5                               |

Use the fitted model to interpolate and obtain expected values for the vapor pressure of Toluene at the following temperatures: 0, 25, 50, 75, 100, and  $125 (^{\circ}C)$ . Since using linear regression to fit the equation to the data will require a variable transformation, obtain 95% confidence intervals for these expected values first in

---

<sup>13</sup>B.A. Ogunnaike and W. H. Ray, (1994). *Process Dynamics, Modeling and Control*, Oxford University Press, NY. Chapter 5.

the transformed variables and convert to approx confidence interval in the original variables.

**16.34** In September 2007, two graduate students<sup>14</sup> studying at the African Institute of Mathematical Sciences (AIMS) in Muizenberg, South Africa, took the following measurements of “wingspan,” (the fingertip-to-fingertip length of outstretched hands) and height for 36 of their classmates.

| “Wingspan”(cm) | Height(cm) | “Wingspan”(cm) | Height(cm) |
|----------------|------------|----------------|------------|
| 182.50         | 171.00     | 165.50         | 158.00     |
| 167.50         | 161.50     | 193.00         | 189.50     |
| 175.00         | 170.00     | 198.00         | 183.00     |
| 163.00         | 164.00     | 181.50         | 181.00     |
| 186.50         | 180.00     | 154.00         | 157.00     |
| 168.50         | 162.50     | 168.00         | 165.00     |
| 166.50         | 158.00     | 174.00         | 166.50     |
| 156.00         | 157.00     | 180.00         | 172.00     |
| 153.00         | 156.00     | 173.00         | 171.50     |
| 170.50         | 162.00     | 188.00         | 179.00     |
| 164.50         | 157.50     | 188.00         | 176.00     |
| 170.50         | 165.50     | 180.00         | 178.00     |
| 173.00         | 164.00     | 160.00         | 163.00     |
| 189.00         | 182.00     | 200.00         | 184.00     |
| 179.50         | 174.00     | 177.00         | 180.00     |
| 174.50         | 165.00     | 179.00         | 169.00     |
| 186.00         | 175.00     | 197.00         | 183.00     |
| 192.00         | 188.00     | 168.50         | 165.00     |

- (i) Obtain a two-parameter model relating height as  $y$  to “wingspan” as  $x$ . Are the two parameters significant? Inspect the residuals and comment on what they imply about the data, the regression model, and how “predictable” height is from a measurement of “wingspan.”
- (ii) If two new students, one short (156.5 cm), one tall (187 cm), arrive in class later in the school year, estimate the respective expected “wingspans” for these new students, along with 95% confidence intervals.

**16.35** Heusner, (1991)<sup>15</sup> compiled a collection of basal metabolism rate (BMR) values,  $Q_0$  (in Watts), and body mass,  $M$  (g), of 391 mammalian species. The table below is a sample from this collection. The relationship between these variables, known as Kleiber’s law, is:

$$Q = \theta_0 M^{\theta_1}$$

From this sample data, use a logarithmic transformation and estimate the unknown parameters, along with 95% confidence intervals. Use the unit of kg for  $M$  in the regression equation.

In the same article, Heusner presented some theoretical arguments for why the exponent should be  $\theta_1 = 2/3$  across species. Compare your estimate with this theo-

<sup>14</sup>Ms. Tabitha Gathoni Mundia from Kenya and Mr. Simon Peter Johnstone-Robertson from South Africa.

<sup>15</sup>Heusner, A.A. (1991). “Size and Power in Mammals,” *J. Exp. Biol.* 160, 25-54.

retical value.

| Species                        | Body Mass<br><i>M</i> (g) | BMR<br><i>Q</i> <sub>0</sub> (Watts) |
|--------------------------------|---------------------------|--------------------------------------|
| <i>Camelus dromedarius</i>     | 407000.00                 | 229.18                               |
| <i>Sus scrofa</i>              | 135000.00                 | 104.15                               |
| <i>Tragulus javanicus</i>      | 1613.00                   | 4.90                                 |
| <i>Ailurus fulgens</i>         | 5740.00                   | 5.11                                 |
| <i>Arctitis binturong</i>      | 14280.00                  | 12.54                                |
| <i>Canis latrans</i>           | 10000.00                  | 14.98                                |
| <i>Herpestes auropunctatus</i> | 611.00                    | 2.27                                 |
| <i>Meles meles</i>             | 11050.00                  | 16.80                                |
| <i>Mustela frenata</i>         | 225.00                    | 1.39                                 |
| <i>Anoura caudifer</i>         | 11.50                     | 0.24                                 |
| <i>Chrotopterus auritus</i>    | 96.10                     | 0.80                                 |
| <i>Eptesicus fuscus</i>        | 16.90                     | 0.11                                 |
| <i>Macroderma gigas</i>        | 148.00                    | 0.78                                 |
| <i>Noctilio leporinus</i>      | 61.00                     | 0.40                                 |
| <i>Myrmecophaga tridactyla</i> | 30600.00                  | 14.65                                |
| <i>Priodontes maximus</i>      | 45190.00                  | 17.05                                |
| <i>Crocidura suaveolens</i>    | 7.50                      | 0.12                                 |
| <i>Didelphis marsupialis</i>   | 1329.00                   | 3.44                                 |
| <i>Lasiorhinus latifrons</i>   | 25000.00                  | 14.08                                |
| <i>Elephas maximus</i>         | 3672 000.00               | 2336.50                              |

# Chapter 17

## Probability Model Validation

|   |     |
|---|-----|
| 17.1 Introduction .....                                   | 732 |
| 17.2 Probability Plots .....                              | 732 |
| 17.2.1 Basic Principles .....                             | 733 |
| 17.2.2 Transformations and Specialized Graph Papers ..... | 734 |
| 17.2.3 Modern Probability Plots .....                     | 736 |
| 17.2.4 Applications .....                                 | 736 |
| Safety Data .....   | 737 |
| Yield Data .....  | 737 |
| Residual Analysis for Regression Model .....              | 737 |
| Others .....  | 737 |
| 17.3 Chi-Squared Goodness-of-fit Test .....               | 739 |
| 17.3.1 Basic Principles .....                             | 739 |
| 17.3.2 Properties and Application .....                   | 741 |
| Poisson Model Validation .....                            | 742 |
| Binomial Special Case .....                               | 743 |
| 17.4 Summary and Conclusions .....                        | 745 |
| REVIEW QUESTIONS .....                                    | 746 |
| EXERCISES .....   | 747 |
| APPLICATION PROBLEMS .....                                | 750 |

*An abstract analysis which is accepted  
without any synthetic examination  
of the question under discussion  
is likely to surprise rather than enlighten us*

Daniel Bernoulli (1700–1782)

In his pithy statement, “All models are wrong but some are useful,” the legendary George E. P. Box of Wisconsin was employing hyperbole to make a subtle but important point. The point, well-known to engineers, is that perfection is not a prerequisite for usefulness in modeling (in fact, it can be an impediment). If complex, real-world problems are to become tractable, idealizing assumptions are inevitable. But what is thus given up in “perfection” is more than made up for in usefulness, *so long as the assumptions can be validated as reasonable*. As a result, assessing the reasonableness of inevitable assumptions is — or ought to be — an important part of the modeling exercise; and this chapter is concerned with presenting some techniques for doing just that — validating distributional assumptions. We focus specifically on probability plots and the chi-squared goodness-of-fit test, two time-tested techniques that also happen to complement each other perfectly in such a way that, with one

or the other, we are able to deal with both discrete and continuous probability models.

## 17.1 Introduction

When confronted with a problem involving a randomly varying phenomenon, the approach we have advocated thus far involves first characterizing the random random phenomenon in question with an ideal probability model in the form of the pdf,  $f(x; \theta)$ , and then using the model to solve the problem at hand. As we have seen in the preceding chapters, fully characterizing the random phenomenon itself involves first (i) postulating a candidate probability model (for example, the Poisson model for the glass inclusions data of Chapter 1) based on an understanding of the underlying phenomenon; and then, (ii) using statistical inference techniques to obtain (point and interval) estimates of the unknown parameter vector  $\theta$  based on sample data,  $X_1, X_2, \dots, X_n$ . However, before proceeding to use the postulated model to solve the problem at hand, it is always advisable to check to be sure that the model and the implied underlying assumptions are reasonable. The question of interest is therefore as follows:

Given sample data,  $X_1, X_2, \dots, X_n$ , obtained from a random variable,  $X$ , with postulated pdf,  $f(x)$ , (and a corresponding cumulative distribution function (cdf),  $F(x)$ ), is the postulated probability model “reasonable”?

The issue of probability model validation is considered in this chapter in its broadest sense: from checking the reasonableness of any postulated probability model for a “free-standing” random variable,  $X$ , to the validation of the near-universal normality assumption for the residuals in regression analysis discussed in Chapter 16. We focus on two approaches: (i) probability plotting and (ii) the chi-squared goodness-of-fit test. The first technique is a classic staple that has evolved from its strictly visual (and hence subjective), old-fashioned probability paper roots, to the more modern, computer-based incarnation, supported with more rigorous statistical analysis. It is better suited to continuous probability models. The second technique is a standard hypothesis test based on a Chi-squared test statistic. While more versatile, being applicable to all random variable types, it is nevertheless applicable *more directly* to discrete probability models.

## 17.2 Probability Plots

### 17.2.1 Basic Principles

Consider that specific experimental data,  $x_1, x_2, \dots, x_n$ , have been obtained from a population whose pdf is postulated as  $f(x)$  (so that the corresponding cdf,  $F(x)$  is also known). To use this data set to check the validity of such a distributional assumption, we start by ordering the data from the smallest to the largest, as  $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ , such that:  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ , i.e.,  $x_{(1)}$  is the smallest of the set, followed by  $x_{(2)}$ , etc., with  $x_{(n)}$  as the largest. For example, one of the data sets on the waiting time (in days) until the occurrence of a recordable safety incident in a certain company's manufacturing site, was given in Example 14.3 as  $S_1 = \{16, 1, 9, 34, 63, 44, 1, 63, 41, 29\}$ , a sample of size  $n = 10$ . This was postulated to be from an exponential distribution. When rank ordered, this data set will be  $S_1^r = \{1, 1, 9, 16, 29, 34, 41, 44, 63, 63\}$ .

Now, observe that because of random variability,  $X_{(i)n}$ , the  $i^{th}$  ranked observation of an  $n$ -sample set, will be a random variable whose observed value will change from one sample to the next. For example, we recall from the same Example 14.3 that a second sample of size 10 obtained from the same company the following year, was given as:  $S_2 = \{35, 26, 16, 23, 54, 13, 100, 1, 30, 31\}$ ; it is also considered to be from the same exponential distribution. When rank ordered, this data set yields:  $S_2^r = \{1, 13, 16, 23, 26, 30, 31, 35, 54, 100\}$ . Note that the value for  $x_{(5)}$ , the fifth ranked from below, is 29 for  $S_1$ , but 26 for  $S_2$ .

Now, define the expected value of  $X_{(i)n}$  as:

$$E(X_{(i)n}) = \mu_{(i)n} \quad (17.1)$$

This quantity can be computed for any given  $f(x)$ , in much the same way that we are able to compute the expectation,  $E(X)$ , of the regular, unranked random variable. The fundamental principle behind probability plots is that if the sample data set truly came from a population with the postulated pdf, then a plot of the ordered sample observations,  $x_{(i)}$ , against their respective expected values,  $\mu_{(i)n}$ , will lie on a straight line, with deviations due only to random variability. Any significant departure from this straight line will indicate that the distributional assumption is not true.

However, with the exception of the very simplest of pdfs, obtaining an exact closed form expression for  $E(X_{(i)n}) = \mu_{(i)n}$  is not a trivial exercise. Nevertheless, using techniques that lie outside the intended scope of this book, it is possible to show that the expression

$$E(X_{(i)n}) = \mu_{(i)n} = F^{-1}\left(\frac{i - \kappa}{n - 2\kappa + 1}\right); i = 1, 2, \dots, n \quad (17.2)$$

is a very good approximation that is valid for all pdfs (the constant  $\kappa$  is defined

later). Here  $F^{-1}(.)$  represents the inverse cumulative distribution function, i.e., the value of  $x$  such that:

$$F(x) = \int_{-\infty}^x f(\xi)d\xi = \left( \frac{i - \kappa}{n - 2\kappa + 1} \right) \quad (17.3)$$

in which case

$$P [X \leq E(X_{(i)n})] = \frac{i - \kappa}{n - 2\kappa + 1} \quad (17.4)$$

For our purposes here, the most important implication of this result is that if  $i$  is the rank of the rank ordered observation,  $x_{(i)}$ , from a population with a postulated pdf  $f(x)$ , then the associated theoretical cumulative probability is  $(i - \kappa)/(n - 2\kappa + 1)$ . In other words, the  $\frac{(i - \kappa)}{(n - 2\kappa + 1)} \times 100\%$  percentile determined from the theoretical cdf is  $E(X_{(i)n})$ .

The constant,  $\kappa$ , depends on sample size,  $n$ , and on  $f(x)$ . However, for all practical purposes, the value  $\kappa = 0.5$  has been found to work quite well for a wide variety of distributions, the exception being the uniform distribution, for which a closed form expression is easily obtained as  $E(X_{(i)n}) = \mu_{(i)n} = i/(n + 1)$ , so that in this case, the appropriate value is  $\kappa = 0$ .

Observe in summary, therefore, that the principle behind the probability plot calls for rank ordering the data, then plotting the rank ordered data,  $x_{(i)}$ , versus its expected value,  $\mu_{(i)}$  (where for convenience, we have dropped the indicator of sample size). From Eq (17.3), obtaining  $\mu_{(i)}$  requires computing the value of the  $(i - \kappa)/(n - 2\kappa + 1)$  quantile from the theoretical cdf,  $F(x)$ . A plot of  $\mu_{(i)}$  on a regular, uniform scale cartesian  $y$ -axis, against  $x_{(i)}$  on the similarly scaled cartesian  $x$ -axis, will show a straight line relationship when the underlying assumptions are reasonable.

### 17.2.2 Transformations and Specialized Graph Papers

The amount of computational effort required in determining  $\mu_{(i)}$  from Eqns (17.2) and (17.3) can be substantial. However, observe from (17.2) that, instead of plotting  $\mu_{(i)}$  on a regular scale, scaling the  $y$ -axis appropriately by  $F(.)$ , the cdf in question, allows us to plot  $x_{(i)}$  directly versus the cumulative probability itself,

$$q_i = \frac{(i - \kappa)}{(n - 2\kappa + 1)} \quad (17.5)$$

a much easier proposition that avoids the need to compute  $E(X_{(i)n})$  first.

This is the fundamental concept behind the old-fashioned probability papers whose scales are tailored for specific cdfs. The most popular of these specialized graph papers is the normal probability paper where the scale is wider at the low end ( $q_i \approx 0$ ), narrowing towards the middle ( $q_i \approx 0.5$ ) and then widening out again symmetrically towards the high end ( $q_i \approx 1$ ). Most of these pre-lined graph sheets were constructed for  $q_i \times 100\%$  (for percentile) on the  $x$ -axis, along with a regular, uniform  $y$ -axis for the rank ordered data. The

**TABLE 17.1:** Table of values for safety data probability plot

| Original Data<br>$x$ | Rank Ordered Data<br>$x_{(i)}$ | Data Rank<br>$i$ | Cumulative Probability<br>$q_i = \left(\frac{i-0.5}{10}\right)$ |
|----------------------|--------------------------------|------------------|---|
| 16                   | 1                              | 1                | 0.05  |
| 1                    | 1                              | 2                | 0.15  |
| 9                    | 9                              | 3                | 0.25  |
| 34                   | 16                             | 4                | 0.35  |
| 63                   | 29                             | 5                | 0.45  |
| 44                   | 34                             | 6                | 0.55  |
| 1                    | 41                             | 7                | 0.65  |
| 63                   | 44                             | 8                | 0.75  |
| 41                   | 63                             | 9                | 0.85  |
| 29                   | 63                             | 10               | 0.95  |

resulting probability plots used to test for normality, are routinely referred to as “normal plots” or “normal probability plots.” Corresponding graph papers exist for the exponential, gamma, and a few other distributions.

The advent of modern computer software packages has not only made these graphs obsolete; it has also made it possible for the technique to be more objective. But before proceeding to discuss the modern approach, we illustrate mechanics of the traditional approach to probability plotting first with the following example.

**Example 17.1: PROBABILITY PLOTTING FOR SAFETY INCIDENT DATA**

Given the data set  $S_1 = \{16, 1, 9, 34, 63, 44, 1, 63, 41, 29\}$  for the waiting time (in days) until the occurrence of a recordable safety incident, generate the data table needed for constructing a probability plot for this putative exponentially distributed random variable. Use  $\kappa = 0.5$ .

**Solution:**

Upon rank-ordering the data, we are able to generate Table 17.1 using  $q_i$  defined as:

$$q_i = \frac{i - 0.5}{10} \quad (17.6)$$

A plot of  $x_{(i)}$  directly versus  $q_i$  on exponential probability paper should yield a straight line if the underlying distribution is truly exponential. Note that without the probability paper, it will be necessary to obtain  $\mu_{(i)}$  first from

$$F(x) = 1 - e^{-\beta x} \quad (17.7)$$

the cdf for the  $E(\beta)$  random variable, i.e.,

$$\mu_{(i)} = -\frac{1}{\beta} \ln(1 - q_i) \quad (17.8)$$

and  $x_{(i)}$  may then be plotted against  $\mu_{(i)}$  on regular graph paper using regular uniform scales on each axis.

With the old-fashioned probability paper, determining how “straight” the plotted data truly were required subjective judgement. The advantage of the visual presentation afforded by the technique therefore had to be tempered by the subjectivity of the “eyeball” assessment. This drawback has now been overcome by the more precise characterization afforded by modern computer programs.

### 17.2.3 Modern Probability Plots

The advent of the computer, and the availability of statistical software packages, has transformed the probability plot from a subjective, purely graphical technique into a much more rigorous and effective probability model validation tool. With programs such as MINITAB, a theoretical distribution line representing the cdf corresponding to the postulated theoretical pdf is obtained, along with a 95% confidence interval, using the probability model parameters estimated from the supplied data (or as provided directly by the user, if available independently). In addition, an Anderson-Darling statistic—a numerical value associated with a test (with the same name) concerning the distribution postulated for the population from which a sample was obtained (see Chapter 18)—is computed along with a  $p$ -value associated with the hypotheses

$$\begin{aligned} H_0 : & \text{ Data follow postulated distribution} \\ H_a : & \text{ Data } do \text{ not } \text{follow postulated distribution} \end{aligned}$$

The displayed probability plot then consists of

1. The rank ordered data, on the  $x$ -axis, versus the cumulative probability (as percentiles) on the appropriately scaled  $y$ -axis, but labeled in the original untransformed values;
2. The theoretical cdf straight line fit along with (approximate) 95% confidence intervals; and
3. A list of basic descriptive statistics and a  $p$ -value associated with the Anderson-Darling (AD) test.

As with other tests, if the  $p$ -value is lower than the chosen  $\alpha$  significance level (typically 0.05), then the null hypothesis is rejected, and we conclude that the data do not follow the hypothesized distribution. In general, we are able to come to one of three conclusions:

1. The model appears to be sufficiently adequate (if  $p > 0.05$ );
2. The model adequacy is indeterminable (if  $0.01 < p < 0.05$ );
3. The model appears inadequate (if  $p < 0.01$ ).

### 17.2.4 Applications

To illustrate probability plotting using MINITAB, we now consider some example data sets encountered in earlier chapters.

#### Safety Data

The probability plots for the safety data sets  $S_1$  and  $S_2$  given above are obtained as follows: after entering the data into respective columns labeled  $X_1$  and  $X_2$ , the sequence: **Graph > Probability Plot**> opens a self-explanatory dialog box; and upon entering all the required information, the plots are generated as shown in Fig 17.1.

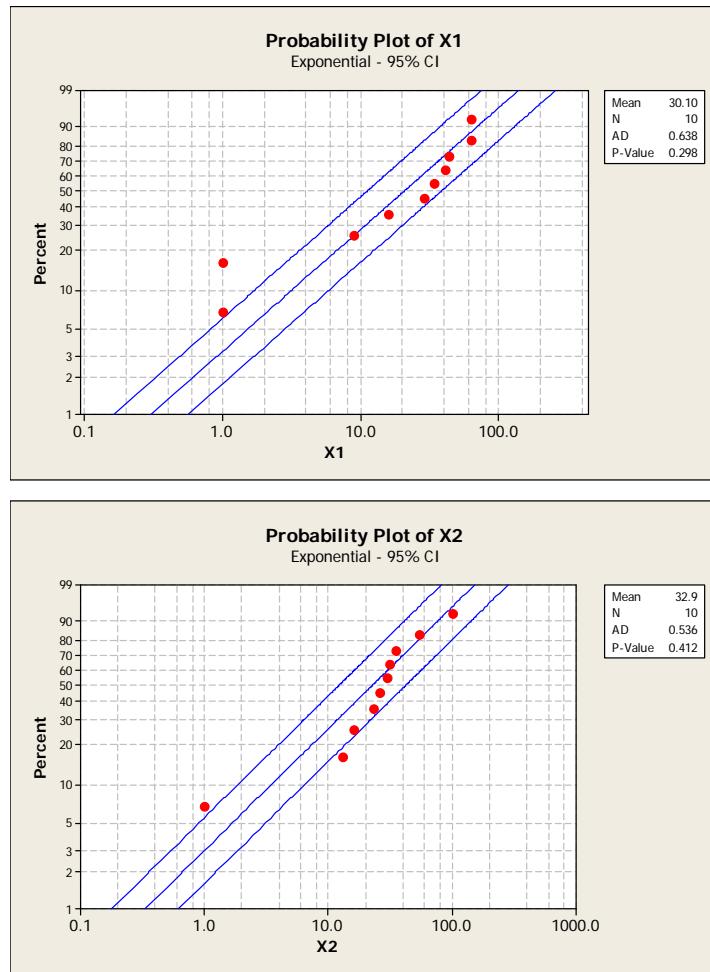
If we had mistakenly postulated the data set  $S_2$  to be normally distributed, for the sake of illustration, the resulting normal probability plot is shown in Fig 17.2. Even though the departure from the straight line fit does not appear to be too severe, the  $p$ -value of 0.045 means that we must reject the null hypothesis at the 0.05 significance level, thereby objectively putting the adequacy of this (clearly wrong) postulate into question.

#### Yield Data

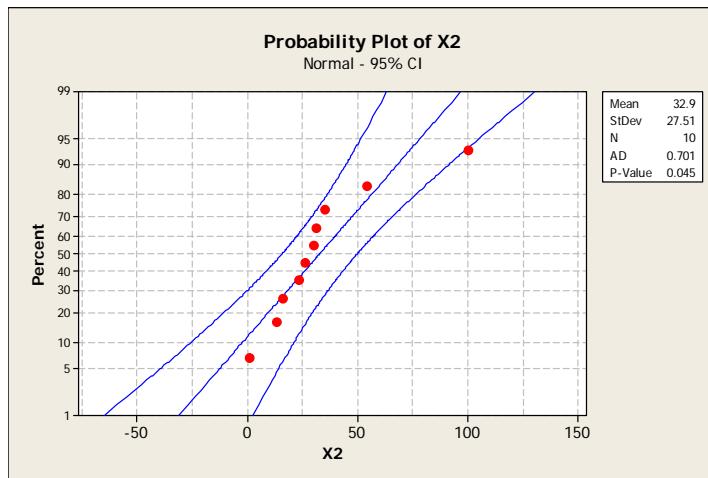
For this illustration, we return to the data sets introduced in Chapter 1 (and subsequently analyzed extensively in ensuing chapters, especially Chapters 14 and 15) on the yields  $Y_A$  and  $Y_B$  obtained from two competing chemical processes  $A$  and  $B$ . Each data set, we may recall, was assumed to have come from Gaussian distributions. The probability plots for these data sets are shown in Fig 17.3. These normal plots, the 95% confidence intervals that completely envelope the entire data sets, and the indicated  $p$ -values, all strongly suggest that these distributional assumptions appear valid.

#### Residual Analysis for Regression Model

The MINITAB regression analysis feature offers as one of its many options, a whole series of residual plots. One of these plots is a “Normal Plot of residuals” which, as the name implies, is a probability plot for the residuals based on the postulate that they are normally distributed. The specific plot shown in Fig 17.4, is obtained directly from within the regression analysis carried out for the data in Example 16.5, where the dependence of thermal conductivity,  $k$ , on Temperature was postulated as a two-parameter regression model with normally distributed random errors. Because it was generated as part of the regression analysis, this plot does not contain the usual additional features of the other probability plots. The residuals can be saved into a data column and separately subjected to the sort of analysis to which the other data sets in this section have been subjected. This is left to the reader as an exercise (See Exercise 17.1).



**FIGURE 17.1:** Probability plots for safety data postulated to be exponentially distributed, each showing (a) rank ordered data; (b) theoretical fitted cumulative probability distribution line along with associated 95% confidence intervals; (c) a list of summary statistics, including the *p*-value associated with a formal goodness-of-fit test. The indication from the *p*-values is that there is no evidence to reject  $H_0$ ; therefore the model appears to be adequate



**FIGURE 17.2:** Probability plot for safety data  $S_2$  wrongly postulated to be normally distributed. The departure from the linear fit does not appear too severe, but the low/borderline  $p$ -value (0.045) objectively compels us to reject  $H_0$  at the 0.05 significance level and conclude that the Gaussian model is inadequate for this data.

### Others

In addition to the probability plots illustrated above for exponential, and Gaussian distributions, MINITAB can also generate probability plots for several other distributions, including lognormal, gamma, and Weibull distributions, all continuous distributions.

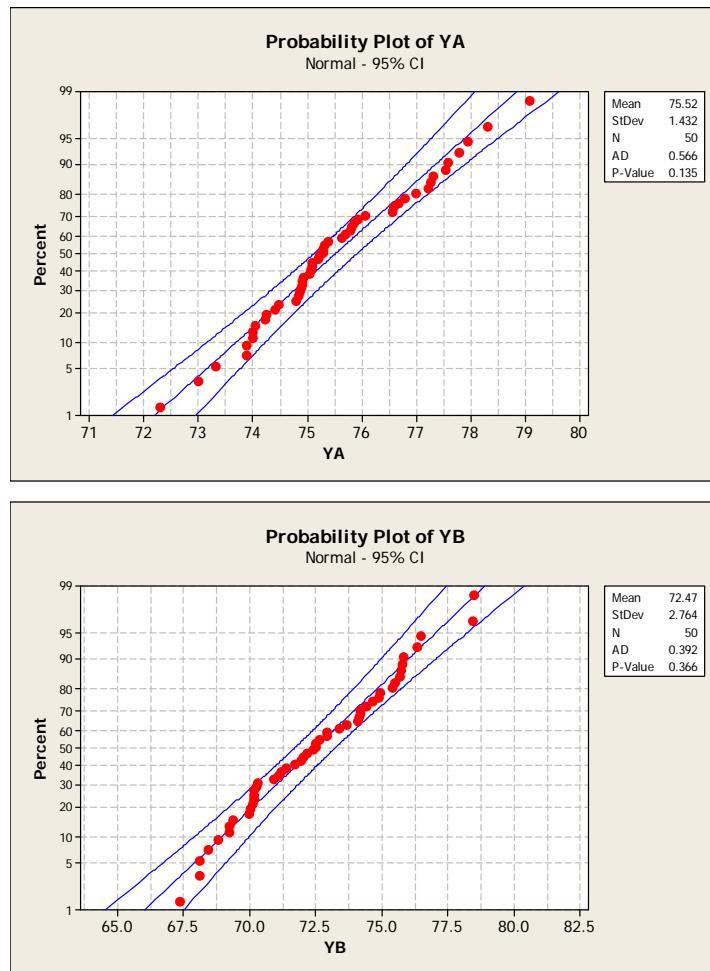
Probability plots are not used for discrete probability models in part because the associated cdfs consist of a series of discontinuous “step” functions, not smooth curves like continuous random variable cdfs. To check the validity of discrete distributions such as the binomial and Poisson, it is necessary to use the more versatile technique discussed next.

## 17.3 Chi-Squared Goodness-of-fit Test

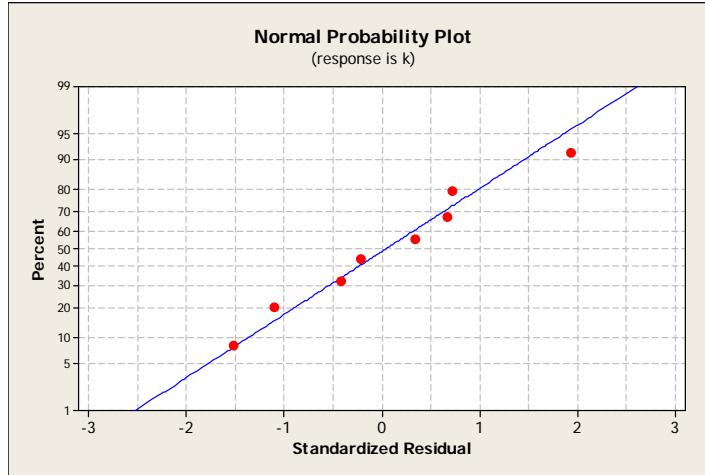
### 17.3.1 Basic Principles

While the probability plot is fundamentally a graphically-based approach, the Chi-squared goodness-of-fit test is fundamentally computational.

We begin by classifying the sample data,  $x_1, x_2, \dots, x_n$ , into  $m$  groups (or “bins”), and obtaining from there a frequency distribution with  $f_i^o$  as the resulting observed frequency associated with the  $i^{th}$  group — precisely



**FIGURE 17.3:** Probability plots for yield data sets  $Y_A$  and  $Y_B$  postulated to be normally distributed. The 95% confidence intervals around the fitted line, along with the indicated  $p$ -values, strongly suggest that the distributional assumptions appear to be valid.



**FIGURE 17.4:** Normal probability plot for the residuals of the regression analysis of the dependence of thermal conductivity,  $k$ , on Temperature in Example 16.5. The postulated model, a two-parameter regression model with Gaussian distributed zero mean errors, appears valid.

how histograms are generated (see Chapter 12). From the postulated probability model, and its  $p$  parameters estimated from the sample data, the theoretical (i.e., expected) frequency associated with each of the  $m$  groups,  $\varphi_i; i = 1, 2, \dots, m$ , is then computed. If the postulated model is correct, the observed and expected frequencies should be close. Because the observed frequencies are subject to random variability, their “closeness” to the corresponding theoretical expectations, quantified by,

$$C^2 = \sum_{i=1}^m \frac{(f_i^o - \varphi_i)^2}{\varphi_i} \quad (17.9)$$

is a statistic that can be shown to have an approximate  $\chi^2(\nu)$  distribution with  $\nu = m - p - 1$  degrees of freedom — an approximation that improves rapidly with increasing  $n$ .

The Chi-squared goodness-of-fit test is a hypothesis test based on this test statistic; the null hypothesis,  $H_0$ , that the data follow the postulated probability model, is tested, at the  $\alpha$  significance level, against the alternative that the data do not follow the model.  $H_0$  is rejected if

$$C^2 > \chi^2_\alpha(\nu) \quad (17.10)$$

### 17.3.2 Properties and Application

The chi-squared goodness-of-fit test is versatile in the sense that it can be applied to both discrete and continuous random variables. With the former, the data already occur naturally in discrete groups; with the latter, theoretical frequencies must be computed by discretizing the continuous intervals. The test is also transparent, logical (as evident from Eq (17.9)) and relatively easy to perform. However, it has some important weaknesses also:

1. To be valid, the test requires that the expected frequency associated with each bin must be at least 5. Where this is not possible, it is recommended that adjacent bins be combined appropriately. This has the drawback that the test will not be very sensitive to tails of postulated models where, by definition, expected observations are few.
2. In general, the test lacks sensitivity in detecting inadequate models when  $n$  is small.
3. Even though recommendations are available for how best to construct discrete intervals for continuous random variables, both the number as well as the nature of these discretized intervals are largely arbitrary and can (and often do) affect the outcome of the test. Therefore, even though applicable in principle, this test is not considered the best option for continuous random variables.

### Poisson Model Validation

The following example illustrates the application of the chi-squared test to the glass manufacturing data presented in Chapter 1 and revisited in various chapters including Chapters 8 (Example 8.8), Chapter 14 (Example 14.13) and Chapter 15 (Example 15.15).

#### **Example 17.2: VALIDATING THE POISSON MODEL FOR INCLUSIONS DATA**

The number of *inclusions* found in each of 60 square-meter sheets of manufactured glass and presented in Table 1.2 in Chapter 1, was postulated to be a Poisson random variable with the single parameter,  $\lambda$ . Perform a chi-squared goodness-of-fit test on this data to evaluate the reasonableness of this postulate.

#### **Solution:**

Recall from our various encounters with this data set that the Poisson model parameter, estimated from the data mean, is  $\hat{\lambda} = 1.017$ . If the data are now arranged into frequency groups for 0, 1, 2, and 3+ inclusions, we obtain the following table:

| Data Group<br>(Inclusions) | Observed<br>Frequency | Poisson<br>$f(x \lambda)$ | Expected<br>Frequency |
|----------------------------|-----------------------|---------------------------|-----------------------|
| 0                          | 22                    | 0.3618                    | 21.708                |
| 1                          | 23                    | 0.3678                    | 22.070                |
| 2                          | 11                    | 0.1870                    | 11.219                |
| $\geq 3$                   | 4                     | 0.0834                    | 5.004                 |

with the expected frequency obtained from  $60 \times f(x|\lambda)$ . We may now compute the desired  $C^2$  statistic as:

$$\begin{aligned} C^2 &= \frac{(22 - 21.708)^2}{21.708} + \frac{(23 - 22.070)^2}{22.070} + \frac{(11 - 11.219)^2}{11.219} + \frac{(4 - 5.004)^2}{5.004} \\ &= 0.249 \end{aligned} \quad (17.11)$$

The associated degrees of freedom is  $4 - 1 - 1 = 2$ , so that from the  $\chi^2(2)$  distribution, we obtain

$$P(\chi^2(2) > 0.249) = 0.883 \quad (17.12)$$

As a result, we have no evidence to reject the null hypothesis, and hence conclude that the Poisson model for this data set appears adequate.

Of course, it is unnecessary to carry out any of the indicated computations by hand, even the frequency grouping. Programs such as MINITAB have chi-squared test features that can be used for problems of this kind.

When MINITAB is used on this last example, upon entering the raw data into a column labeled “Inclusions,” the sequence, **Stat > Basic Stats > Chi-Sq Goodness-of-Fit-Test for Poisson** opens a self-explanatory dialog box; and making the required selections produces the following results, just as we had obtained earlier:

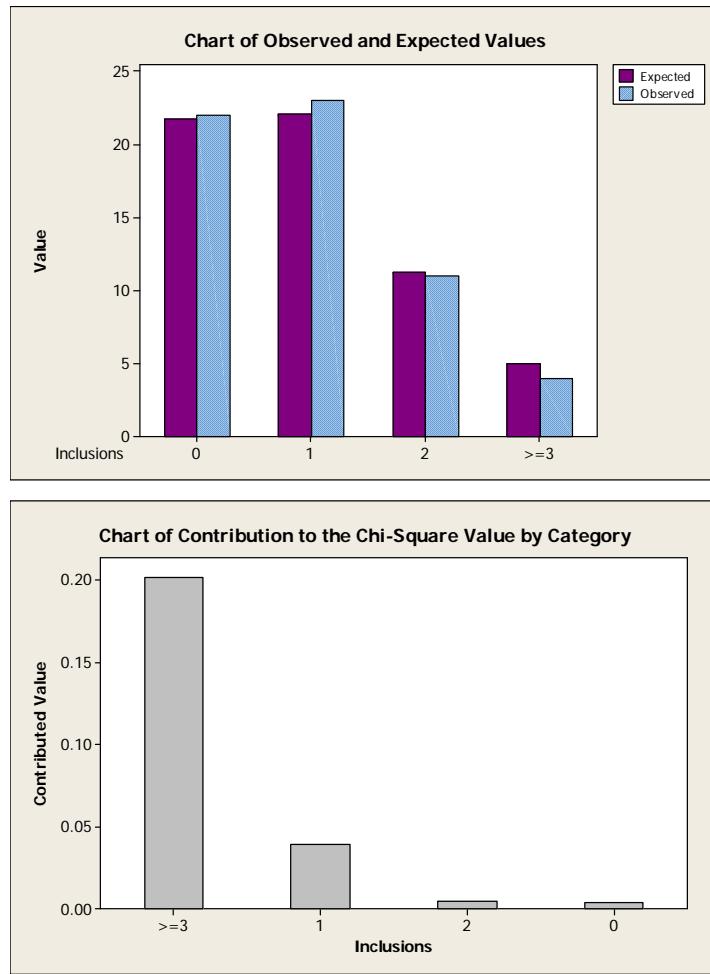
#### Goodness-of-Fit Test for Poisson Distribution

Data column: Inclusions

Poisson mean for Inclusions = 1.01667

| Inclusions | Observed | Poisson     |          | Contribution<br>to Chi-Sq |
|------------|----------|-------------|----------|---------------------------|
|            |          | Probability | Expected |                           |
| 0          | 22       | 0.361799    | 21.7079  | 0.003930                  |
| 1          | 23       | 0.367829    | 22.0697  | 0.039212                  |
| 2          | 11       | 0.186980    | 11.2188  | 0.004267                  |
| $\geq 3$   | 4        | 0.083392    | 5.0035   | 0.201279                  |
|            |          | N           | DF       | Chi-Sq P-Value            |
|            |          | 60          | 2        | 0.248687 0.883            |

The rightmost column in the MINITAB output shows the individual contributions from each group to the chi-squared statistic — an indication of how the “lack-of-fit” is distributed among the groups. For example, the group of 3 or more inclusions contributed by far the largest to the discrepancy between observation and model prediction; but even this is not sufficient to jeopardize the model adequacy. MINITAB also produces graphical representations of these results, as shown in Fig 17.5.



**FIGURE 17.5:** Chi-Squared test results for inclusions data and a postulated Poisson model. Top panel: Bar chart of “Expected” and “Observed” frequencies, which shows how well the model prediction matches observed data; Bottom Panel: Bar chart of contributions to the Chi-squared statistic, showing that the group of 3 or more inclusions is responsible for the largest model-observation discrepancy, by a wide margin.

### Binomial Special Case

For the binomial case, where there are only two categories ( $x$  “successes” and  $n - x$  “failures” observed in  $n$  independent trials being the observed frequencies in each respective category), for a postulated  $Bi(n, p)$  model, the chi-squared statistic reduces to:

$$C^2 = \frac{(x - np)^2}{np} + \frac{[(n - x) - nq]^2}{nq} \quad (17.13)$$

where  $q = 1 - p$ , as usual. When this expression is consolidated to

$$C^2 = \frac{q(x - np)^2 + p[(n - x) - nq]^2}{npq}$$

upon introducing  $q = 1 - p$  for the first term in the numerator and taking advantage of the “difference of two squares” result in algebra, the right hand side of the equation rearranges easily to give the result:

$$C^2 = \frac{(x - np)^2}{npq} \quad (17.14)$$

which, if we take the positive square root, reduces to:

$$Z = \frac{x - np}{\sqrt{npq}} \quad (17.15)$$

This, of course, is immediately recognized as the  $z$ -statistic for the (large sample) Gaussian approximation to the binomial random variable used to carry out the  $z$ -test of the observed mean against a postulated mean  $np$ , as discussed in Chapter 15. Thus, the chi-squared test for the binomial model is identical to the standard  $z$ -test when the population parameter  $p$  is specified independently.

### 17.4 Summary and Conclusions

This chapter has been primarily concerned with examining two methods for validating probability models: modern probability plots and the chi-square goodness-of-fit test. While we presented the principles behind these methods, we concentrated more on applying them, particularly with the aid of computer programs. With some perspective, we may now observe the following as the main points of the chapter:

- Probability plots augmented with theoretical model fits and  $p$ -values are most appropriate for continuous models;

- Chi-squared tests, on the other hand, are more naturally suited to discrete models (although they can also be applied to continuous models after appropriate discretization).

As a practical matter, it is important to keep in mind that, just as with other hypotheses tests, a postulated probability model can never be completely *proven* adequate by these tests (on the basis of finite sample data), but inadequate models can be successfully identified as such. Still, it can be difficult to identify inadequate models with these tests when sample sizes are small; our chances of identifying inadequate models correctly as inadequate improve significantly as  $n \rightarrow \infty$ . Therefore, as much sample data as possible should be used to validate probability models; and wherever possible, the data set used to validate a model should be collected independently of that used to estimate the parameters. Some of the end-of-chapter exercises and applications problems are used to reinforce these points.

Finally, it must be kept in mind always that no model is (or can ever be) perfect. The final decision about the validity of the model assumptions rests with the practitioner—the person who will ultimately use these models for problem solving—and these tests should be considered properly only as objective guides, not as final and absolute arbiters.

---

## REVIEW QUESTIONS

1. What is the primary question of interest in probability model validation?
2. What are the two approaches discussed in this chapter for validating probability models?
3. Which approach is better suited to continuous probability models and which one is applicable most directly to discrete probability models?
4. What is the fundamental principle behind probability plots?
5. What is the fundamental concept behind old-fashioned probability plots?
6. What hypothesis test accompanies modern probability plots?
7. What does a modern probability plot consist of?
8. Why are probability plots not used for discrete probability models?
9. What is a chi-squared goodness-of-fit test?

10. What are some weaknesses of the chi-squared goodness-of-fit test?
11. The chi-squared goodness-of-fit test for the binomial model is identical to what familiar hypothesis test?

---

## EXERCISES

17.1 In Example 16.5, a two-parameter regression model of how a metal's thermal conductivity varies with temperature was developed from the data shown here again for ease of reference.

| $k$ (W/m-°C) | Temperature (°C) |
|--------------|------------------|
| 93.228       | 100              |
| 92.563       | 150              |
| 99.409       | 200              |
| 101.590      | 250              |
| 111.535      | 300              |
| 115.874      | 350              |
| 119.390      | 400              |
| 126.615      | 450              |

The two-parameter model was postulated for the relationship between  $k$  and  $T$  with the implicit assumption that the errors are normally distributed. Obtain the residuals from the least-squares fit and generate a normal probability plot (and ancillary analysis) of these residuals. Comment on the validity of the normality assumption for the regression model errors.

17.2 In Problem 16.28, the data in the table below<sup>1</sup> was presented as the basis for calibrating a near-infrared instrument to be used to determine protein content in wheat from reflectance measurements.

For this data set to produce a useful calibration curve, the regression model must be adequate; and an important aspect of regression model adequacy is the nature of its residuals. In this particular case, the residuals are required to be random and approximately normally distributed. By analyzing the residuals from the regression exercise appropriately, comment on whether or not the resulting regression model should be considered as adequate.

---

<sup>1</sup>Fern, A.T., (1983). "A misuse of ridge regression in the calibration of a near-infrared reflectance instrument," *Applied Statistics*, 32, 73–79

| Protein Content %<br><i>y</i> | Observed Reflectance<br><i>x</i> | Protein Content %<br><i>y</i> | Observed Reflectance<br><i>x</i> |
|-------------------------------|----------------------------------|-------------------------------|----------------------------------|
| 9.23                          | 386                              | 10.57                         | 443                              |
| 8.01                          | 383                              | 10.23                         | 450                              |
| 10.95                         | 353                              | 11.87                         | 467                              |
| 11.67                         | 340                              | 8.09                          | 451                              |
| 10.41                         | 371                              | 12.55                         | 524                              |
| 9.51                          | 433                              | 8.38                          | 407                              |
| 8.67                          | 377                              | 9.64                          | 374                              |
| 7.75                          | 353                              | 11.35                         | 391                              |
| 8.05                          | 377                              | 9.70                          | 353                              |
| 11.39                         | 398                              | 10.75                         | 445                              |
| 9.95                          | 378                              | 10.75                         | 383                              |
| 8.25                          | 365                              | 11.47                         | 404                              |

**17.3** The following data is postulated to have been sampled from an exponential  $\mathcal{E}(4)$  population. Validate the postulated model appropriately. Repeat the validation exercise as if the population parameter was unknown and hence must be estimated from the sample data. Does knowing the population parameter independently make any difference in this particular case?

|       |      |      |      |       |
|-------|------|------|------|-------|
| 6.99  | 2.84 | 0.41 | 3.75 | 2.16  |
| 0.52  | 0.67 | 2.72 | 5.22 | 16.65 |
| 10.36 | 1.66 | 3.26 | 1.78 | 1.31  |
| 5.75  | 0.12 | 6.51 | 4.05 | 1.52  |

**17.4** The data below are random samples from two independent lognormal distributions; specifically,  $X_{L_1} \sim \mathcal{L}(0, 0.25)$  and  $X_{L_2} \sim \mathcal{L}(0.25, 0.25)$ .

| $X_{L_1}$ | $X_{L_2}$ |
|-----------|-----------|
| 0.81693   | 1.61889   |
| 0.96201   | 1.15897   |
| 1.03327   | 1.17163   |
| 0.84046   | 1.09065   |
| 1.06731   | 1.27686   |
| 1.34118   | 0.91838   |
| 0.77619   | 1.45123   |
| 1.14027   | 1.47800   |
| 1.27021   | 2.16068   |
| 1.69466   | 1.46116   |

- (i) Test the validity of these statements *directly* from the data as presented.
- (ii) Test the validity of these statements *indirectly* by taking a logarithmic transformation of the data, and carrying out an appropriate analysis of the resulting log-transformed data. Compare the results with those obtained in (i).

**17.5** If  $X_1$  is a lognormal random variable with parameters  $(\alpha, \beta_1)$ , and  $X_2$  is a lognormal random variable with parameters  $(\alpha, \beta_2)$ , it has been postulated that the product:

$$Y = X_1 X_2$$

has a lognormal distribution with parameters  $(\alpha, \beta_y)$  where:

$$\beta_y = \beta_1 + \beta_2$$

- (i) Confirm this result theoretically. (*Hint:* use results from Chapter 6 regarding the sum of Gaussian random variables.)
- (ii) In the data table below,  $X_1$  is a random sample drawn from a distribution purported to be  $\mathcal{L}(0.25, 0.50)$ ; and  $X_2$  is a random sample drawn from a distribution purported to be  $\mathcal{L}(0.25, 0.25)$ .

| $X_1$   | $X_2$   |
|---------|---------|
| 1.16741 | 1.61889 |
| 1.58631 | 1.15897 |
| 2.00530 | 1.17163 |
| 1.67186 | 1.09065 |
| 1.63146 | 1.27686 |
| 1.61738 | 0.91838 |
| 0.74154 | 1.45123 |
| 2.96673 | 1.47800 |
| 1.50267 | 2.16068 |
| 1.99272 | 1.46116 |

From this data set, obtain the corresponding values for  $Y$  defined as the product  $Y = X_1 X_2$ . According to the result stated and proved in (i), what is the theoretical distribution of  $Y$ ? Confirm that the computed sample data set for  $Y$  agrees with this postulate.

**17.6** The data in the following table (Exercise 12.12) shows samples of size  $n = 20$  drawn from four different populations postulated to be normal,  $N$ , lognormal  $L$ , gamma  $G$ , and inverse gamma  $I$ , respectively.

| $X_N$   | $X_L$   | $X_G$   | $X_I$    |
|---------|---------|---------|----------|
| 9.3745  | 7.9128  | 10.0896 | 0.084029 |
| 8.8632  | 5.9166  | 15.7336 | 0.174586 |
| 11.4943 | 4.5327  | 15.0422 | 0.130492 |
| 9.5733  | 33.2631 | 5.5482  | 0.115567 |
| 9.1542  | 24.1327 | 18.0393 | 0.187260 |
| 9.0992  | 5.4151  | 17.9543 | 0.100054 |
| 10.2631 | 16.9556 | 12.5549 | 0.101405 |
| 9.8737  | 3.9345  | 9.6640  | 0.100835 |
| 7.8192  | 35.0376 | 14.2975 | 0.097173 |
| 10.4691 | 25.1182 | 4.2599  | 0.141233 |
| 9.6981  | 1.1804  | 19.1084 | 0.060470 |
| 10.5911 | 2.3503  | 7.0735  | 0.127663 |
| 11.6526 | 15.6894 | 7.6392  | 0.074183 |
| 10.4502 | 5.8929  | 14.1899 | 0.086606 |
| 10.0772 | 8.0254  | 13.8996 | 0.084915 |
| 10.2932 | 16.1482 | 9.7680  | 0.242657 |
| 11.7755 | 0.6848  | 8.5779  | 0.052291 |
| 9.3790  | 6.6974  | 7.5486  | 0.116172 |
| 9.9202  | 3.6909  | 10.4043 | 0.084339 |
| 10.9067 | 34.2152 | 14.8254 | 0.205748 |

- (i) Validate these postulates using the full data sets. Note that the population parameters have not been specified.  
(ii) Using only the top half of each data set, repeat (i). For this particular example, what effect, if any, does sample size have on the probability plots approach to probability model validation?

**17.7** The data in the table below was presented in Exercise 15.18 as a random sample of 15 observations each from two normal populations with unknown means and variances. Test the validity of the normality assumption for each data set. Interpret your results.

| Sample | X     | Y     |
|--------|-------|-------|
| 1      | 12.03 | 13.74 |
| 2      | 13.01 | 13.59 |
| 3      | 9.75  | 10.75 |
| 4      | 11.03 | 12.95 |
| 5      | 5.81  | 7.12  |
| 6      | 9.28  | 11.38 |
| 7      | 7.63  | 8.69  |
| 8      | 5.70  | 6.39  |
| 9      | 11.75 | 12.01 |
| 10     | 6.28  | 7.15  |
| 11     | 12.53 | 13.47 |
| 12     | 10.22 | 11.57 |
| 13     | 7.17  | 8.81  |
| 14     | 11.36 | 13.10 |
| 15     | 9.16  | 11.32 |

## APPLICATION PROBLEMS

**17.8** The following data set, from a study by Lucas (1985)<sup>2</sup>, shows the number of accidents occurring per quarter (three months) at a DuPont company facility, over a 10-year period. The data set has been partitioned into two periods: Period I is the first five-year period of the study; Period II, the second five-year period.

| Period I |   |    |    | Period II |   |   |   |
|----------|---|----|----|-----------|---|---|---|
| 5        | 5 | 10 | 8  | 3         | 4 | 2 | 0 |
| 4        | 5 | 7  | 3  | 1         | 3 | 2 | 2 |
| 2        | 8 | 6  | 9  | 7         | 7 | 1 | 4 |
| 5        | 6 | 5  | 10 | 1         | 2 | 2 | 1 |
| 6        | 3 | 3  | 10 | 4         | 4 | 4 | 4 |

When the data set was presented in Problem 13.22, it was simply stated as a matter of fact that a Poisson pdf was a reasonable model for representing this data.

<sup>2</sup>Lucas J. M., (1985). "Counted Data CUSUMs," *Technometrics*, 27, 129–144

Check the validity of this assumption for Period I data and for Period II data separately.

**17.9** Refer to Problem 17.8. One means by which one can test if two samples are truly from the same population is to compare the empirical distributions obtained from each data set against each other directly; if the two samples are truly from the same distribution, within the limits of experimental error, there should be no significant difference between the two empirical distributions. State an appropriate null hypothesis and the alternative hypothesis and carry out a chi-squared goodness-of-fit test of the empirical distribution of the data for Period I versus that of Period II. State your conclusions clearly.

**17.10** The table below (see Problem 9.40) shows frequency data on distances between DNA replication origins (inter-origin distances), measured *in vivo* in Chinese Hamster Ovary (CHO) cells by Li et al., (2003)<sup>3</sup>, as reported in Chapter 7 of Birtwistle (2008)<sup>4</sup>. Phenomenologically, inter-origin distance should be a gamma distributed random variable and this data set has been analyzed in Birtwistle, (2008), on this basis. Carry out a formal test to validate the gamma model assumption. Interpret your results.

| Inter-Origin<br>Distance (kb) | Relative<br>Frequency<br>$f_r(x)$ |
|-------------------------------|-----------------------------------|
| 0                             | 0.00                              |
| 15                            | 0.02                              |
| 30                            | 0.20                              |
| 45                            | 0.32                              |
| 60                            | 0.16                              |
| 75                            | 0.08                              |
| 90                            | 0.11                              |
| 105                           | 0.03                              |
| 120                           | 0.02                              |
| 135                           | 0.01                              |
| 150                           | 0.00                              |
| 165                           | 0.01                              |

**17.11** The time in months between occurrences of safety violations in a toll manufacturing facility is shown in the table below for three operators, "A," "B," "C".

|          |      |      |      |      |      |      |      |      |      |      |
|----------|------|------|------|------|------|------|------|------|------|------|
| <i>A</i> | 1.31 | 0.15 | 3.02 | 3.17 | 4.84 | 0.71 | 0.70 | 1.41 | 2.68 | 0.68 |
| <i>B</i> | 1.94 | 3.21 | 2.91 | 1.66 | 1.51 | 0.30 | 0.05 | 1.62 | 6.75 | 1.29 |
| <i>C</i> | 0.79 | 1.22 | 0.65 | 3.90 | 0.18 | 0.57 | 7.26 | 0.43 | 0.96 | 3.76 |

It is customary to postulate an exponential probability model for this phenomenon. Is this a reasonable postulate for each data set in this collection? Support your answer adequately.

<sup>3</sup>Li, F., Chen, J., Solessio, E. and Gilbert, D. M. (2003). "Spatial distribution and specification of mammalian replication origins during G1 phase." *J Cell Biol* 161, 257-66.

<sup>4</sup>M. R. Birtwistle, (2008). *Modeling and Analysis of the ErbB Signaling Network: From Single Cells to Tumorigenesis*, PhD Dissertation, University of Delaware.

**17.12** The data table below (also presented in Problem 8.26) shows  $x$ , a count of the number of species,  $x = 1, 2, \dots, 24$ , and the associated number of Malayan butterflies that have  $x$  number of species. When the data was first published and analyzed in Fisher *et al.*, (1943)<sup>5</sup>, the *logarithmic series distribution* (see Exercise 8.13), with the pdf,

$$f(x) = \frac{\alpha p^x}{x}; 0 < p < 1; x = 1, 2, \dots,$$

where

$$\alpha = \frac{-1}{\ln(1-p)}$$

was proposed as the appropriate model for the phenomenon in question. This pdf has since become the model of choice for data involving this phenomenon.

| $x$                | 1   | 2  | 3  | 4  | 5  | 6  | 7  | 8  |
|--------------------|-----|----|----|----|----|----|----|----|
| No of species      |     |    |    |    |    |    |    |    |
| $\Phi(x)$          | 118 | 74 | 44 | 24 | 29 | 22 | 20 | 19 |
| Observed Frequency |     |    |    |    |    |    |    |    |
| $x$                | 9   | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| No of species      |     |    |    |    |    |    |    |    |
| $\Phi(x)$          | 20  | 15 | 12 | 14 | 6  | 12 | 6  | 9  |
| Observed Frequency |     |    |    |    |    |    |    |    |
| $x$                | 17  | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| No of species      |     |    |    |    |    |    |    |    |
| $\Phi(x)$          | 9   | 6  | 10 | 10 | 11 | 5  | 3  | 3  |
| Observed Frequency |     |    |    |    |    |    |    |    |

Formally validate this model for this specific data set. What is the  $p$ -value associated with the test? What does it indicate about the validity of this model?

**17.13** The data in the table below is the time-to-publication of 85 papers published in the January 2004 issue of a leading chemical engineering research journal. (See Problem 1.13).

---

<sup>5</sup>Fisher, R. A., S. Corbet, and C. B. Williams. (1943). "The relation between the number of species and the number of individuals in a random sample of an animal population." *Journal of Animal Ecology*, 1943: 4258.

|      |      |      |      |      |
|------|------|------|------|------|
| 19.2 | 15.1 | 9.6  | 4.2  | 5.4  |
| 9.0  | 5.3  | 12.9 | 4.2  | 15.2 |
| 17.2 | 12.0 | 17.3 | 7.8  | 8.0  |
| 8.2  | 3.0  | 6.0  | 9.5  | 11.7 |
| 4.5  | 18.5 | 24.3 | 3.9  | 17.2 |
| 13.5 | 5.8  | 21.3 | 8.7  | 4.0  |
| 20.7 | 6.8  | 19.3 | 5.9  | 3.8  |
| 7.9  | 14.5 | 2.5  | 5.3  | 7.4  |
| 19.5 | 3.3  | 9.1  | 1.8  | 5.3  |
| 8.8  | 11.1 | 8.1  | 10.1 | 10.6 |
| 18.7 | 16.4 | 9.8  | 10.0 | 15.2 |
| 7.4  | 7.3  | 15.4 | 18.7 | 11.5 |
| 9.7  | 7.4  | 15.7 | 5.6  | 5.9  |
| 13.7 | 7.3  | 8.2  | 3.3  | 20.1 |
| 8.1  | 5.2  | 8.8  | 7.3  | 12.2 |
| 8.4  | 10.2 | 7.2  | 11.3 | 12.0 |
| 10.8 | 3.1  | 12.8 | 2.9  | 8.8  |

Given the nature of the phenomenon in question, postulate an appropriate model and validate it. If your first postulate is invalid, postulate another one until you obtain a valid model.

**17.14** Refer to Problem 17.13. This time, obtain a histogram and represent the data in the form of a frequency table. Using this “discretized” version of the data, perform an appropriate test to validate the model proposed and validated in Problem 17.13. Is there a difference between the results obtained here and the ones obtained in Problem 17.13? If yes, offer an explanation for what may be responsible for this difference.

**17.15** The distribution of income of families in the US in 1979 (in actual dollars uncorrected for inflation) is shown in the table below (Prob 4.28):

| Income level, $x$ ,<br>$(\times \$10^3)$ | Percent of Population<br>with income level, $x$ |
|--|---|
| 0–5                                      | 4   |
| 5–10                                     | 13  |
| 10–15                                    | 17  |
| 15–20                                    | 20  |
| 20–25                                    | 16  |
| 25–30                                    | 12  |
| 30–35                                    | 7   |
| 35–40                                    | 4   |
| 40–45                                    | 3   |
| 45–50                                    | 2   |
| 50–55                                    | 1   |
| > 55                                     | 1   |

It has been postulated that the lognormal distribution is a reasonable model for this phenomenon. Carry out an appropriate test to confirm or refute this postulate. Keep in mind that the data is not in raw form, but has already been “processed” into a frequency table. Interpret your results.

**17.16** The appropriate analysis of over-dispersed Poisson phenomena with the negative binomial distribution was pioneered with the classic data and analysis of Greenwood and Yule (1920) data<sup>6</sup>. The data in question, shown in the table below (see Problem 8.28), is the frequency of accidents occurring, over a five-week period, to 647 women making high explosives during World War I.

| Number<br>of Accidents | Observed<br>Frequency |
|------------------------|-----------------------|
| 0                      | 447                   |
| 1                      | 132                   |
| 2                      | 42                    |
| 3                      | 21                    |
| 4                      | 3                     |
| 5+                     | 2                     |

First, determine from the data, parameter estimates for a Poisson model and then determine  $k$  and  $p$  for a negative binomial model (see Problem 14.40). Next, conduct formal chi-squared goodness-of-fit tests for the Poisson model and then for the negative binomial. Interpret your test results. From your analysis, which model is more appropriate for this data set?

**17.17** Mee (1990)<sup>7</sup> presented the following data on the wall thickness (in ins) of cast aluminum cylinder heads used in aircraft engine cooling jackets. When presented in Problem 14.44 (and following the maximum entropy arguments of Problem 10.15), the data was assumed to be a random sample from a normal population. Validate this normality assumption and comment on whether or not this is a reasonable assumption.

|       |       |       |       |       |       |       |       |       |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 0.223 | 0.228 | 0.214 | 0.193 | 0.223 | 0.213 | 0.218 | 0.215 | 0.233 |
| 0.201 | 0.223 | 0.224 | 0.231 | 0.237 | 0.217 | 0.204 | 0.226 | 0.219 |

**17.18** A sample of 20 silicon wafers selected and examined for flaws produced the result (the number of flaws found on each wafer) shown in the following table.

|   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|
| 3 | 0 | 0 | 2 | 3 | 0 | 3 | 2 | 1 | 2 |
| 4 | 1 | 2 | 3 | 2 | 1 | 2 | 4 | 0 | 1 |

When this data set was first presented in Problem 12.20, it was suggested that the Poisson model is reasonable for problems of this type. For this particular problem, however, is this a reasonable model? Interpret your results.

**17.19** According to census records, the age distribution of the inhabitants of the United States in 1960 and in 1980 is as shown in the table below.

<sup>6</sup>Greenwood M. and Yule, G. U. (1920) "An enquiry into the nature of frequency distributions representative of multiple happenings with particular reference of multiple attacks of disease or of repeated accidents." *Journal Royal Statistical Society* 83:255-279.

<sup>7</sup>Mee, R. W., (1990). "An improved procedure for screening based on a correlated, normally distributed variable," *Technometrics*, 32, 331-337.

| Age Group | 1960   | 1980   |
|-----------|--------|--------|
| < 5       | 20,321 | 16,348 |
| 5–9       | 18,692 | 16,700 |
| 10–14     | 16,773 | 18,242 |
| 15–19     | 13,219 | 21,168 |
| 20–24     | 10,801 | 21,319 |
| 25–29     | 10,869 | 19,521 |
| 30–34     | 11,949 | 17,561 |
| 35–39     | 12,481 | 13,965 |
| 40–44     | 11,600 | 11,669 |
| 45–49     | 10,879 | 11,090 |
| 50–54     | 9,606  | 11,710 |
| 55–59     | 8,430  | 11,615 |
| 60–64     | 7,142  | 10,088 |
| ≥ 65      | 16,560 | 25,550 |

(i) It is typical to assume that such data are normally distributed. Is this a reasonable assumption in each case? (ii) Visually, the two distributions appear different. But are they significantly so? Carry out an appropriate test to check the validity of any assumption of equality of these two age distributions.

**17.20** In Problem 13.34 and in Example 15.1, it was assumed that the following data, two sets of random samples of trainee scores from large groups of trainees instructed by Method A and Method B, are both normally distributed.

|          |    |    |    |    |    |    |    |    |    |    |
|----------|----|----|----|----|----|----|----|----|----|----|
| Method A | 71 | 75 | 65 | 69 | 73 | 66 | 68 | 71 | 74 | 68 |
| Method B | 72 | 77 | 84 | 78 | 69 | 70 | 77 | 73 | 65 | 75 |

Carry out an appropriate test and confirm whether or not such an assumption is justified.

**17.21** The data below is the computed fractional intensity,  $\phi = I_{test}/(I_{test} + I_{ref})$ , for a collection of special genes (known as “housekeeping genes”), where  $I_{test}$  is the measured fluorescence intensity under test conditions, and  $I_{ref}$ , the intensity under reference conditions. If these 10 genes are true housekeeping genes, then within the limits of measurement noise, the computed values of  $\phi$  should come from a symmetric Beta distribution with mean value 0.5. Use the method of moments to estimate parameter values for the postulated Beta  $B(\alpha, \beta)$  distribution. Carry out an appropriate test to validate the Beta model hypothesis.

| $\phi_i$ |
|----------|
| 0.585978 |
| 0.504057 |
| 0.182831 |
| 0.426575 |
| 0.455191 |
| 0.804720 |
| 0.741598 |
| 0.332909 |
| 0.532131 |
| 0.610620 |

**17.22** Padgett and Spurrier (1990)<sup>8</sup> obtained the following data set for the breaking strengths (in GPa) of carbon fibers used in making composite materials.

|     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 1.4 | 3.7 | 3.0 | 1.4 | 1.0 | 2.8 | 4.9 | 3.7 | 1.8 | 1.6 |
| 3.2 | 1.6 | 0.8 | 5.6 | 1.7 | 1.6 | 2.0 | 1.2 | 1.1 | 1.7 |
| 2.2 | 1.2 | 5.1 | 2.5 | 1.2 | 3.5 | 2.2 | 1.7 | 1.3 | 4.4 |
| 1.8 | 0.4 | 3.7 | 2.5 | 0.9 | 1.6 | 2.8 | 4.7 | 2.0 | 1.8 |
| 1.6 | 1.1 | 2.0 | 1.6 | 2.1 | 1.9 | 2.9 | 2.8 | 2.1 | 3.7 |

- (i) In their analysis, Padgett and Spurrier postulated a Weibull  $W(\zeta, \beta)$  distribution model with parameters  $\zeta = 2.0$  and  $\beta = 2.5$  for the phenomenon in question. Validate this model assumption by carrying out an appropriate test.
- (ii) Had the model parameters not been given, so that their values must be determined from the data, repeat the test in (i) and compare the results. What does this imply about the importance of obtaining independent parameter estimates before carrying out probability model validation tests?

**17.23** The data set below, from Holmes and Mergen (1992)<sup>9</sup>, is a sample of viscosity measurements taken from ten consecutive, but independent, batches of a product made in a batch chemical process.

$$S_{10} = \{13.3, 14.5, 15.3, 15.3, 14.3, 14.8, 15.2, 14.9, 14.6, 14.1\}$$

Part of the assumption in the application noted in the reference is that this data constitutes a random sample from a normal population with unknown mean and unknown variance. Confirm whether or not this is a reasonable assumption.

<sup>8</sup>Padgett, W.J. and J. D. Spurrier, (1990). Shewhart-type charts for percentiles of strength distributions. *J of Quality Tech.* 22, 283–388.

<sup>9</sup>Holmes, D.S., and A.E. Mergen, (1992). “Parabolic control limits for the exponentially weighted moving average control charts” *Qual. Eng.* 4, 487–495.

# Chapter 18

## Nonparametric Methods

|        |  |     |
|--------|--|-----|
| 18.1   | Introduction .....                               | 758 |
| 18.2   | Single Population .....                          | 760 |
| 18.2.1 | One-Sample Sign Test .....                       | 760 |
|        | Basic Test Characteristics .....                 | 760 |
|        | Comparison with Parametric Alternatives .....    | 763 |
| 18.2.2 | One-Sample Wilcoxon Signed Rank Test .....       | 763 |
|        | Basic Test Characteristics .....                 | 763 |
|        | Comparison with Parametric Alternatives .....    | 765 |
| 18.3   | Two Populations .....                            | 765 |
| 18.3.1 | Two-Sample Paired Test .....                     | 766 |
| 18.3.2 | Mann-Whitney-Wilcoxon Test .....                 | 766 |
|        | Basic Test Characteristics .....                 | 766 |
|        | Comparison with Parametric Alternatives .....    | 769 |
| 18.4   | Probability Model Validation .....               | 769 |
| 18.4.1 | The Kolmogorov-Smirnov Test .....                | 770 |
|        | Basic Test Characteristics .....                 | 770 |
|        | Key Features .....                               | 771 |
| 18.4.2 | The Anderson-Darling Test .....                  | 771 |
|        | Key Features .....                               | 772 |
| 18.5   | A Comprehensive Illustrative Example .....       | 772 |
| 18.5.1 | Probability Model Postulate and Validation ..... | 772 |
| 18.5.2 | Mann-Whitney-Wilcoxon Test .....                 | 775 |
| 18.6   | Summary and Conclusions .....                    | 775 |
|        | REVIEW QUESTIONS .....                           | 778 |
|        | EXERCISES .....                                  | 781 |
|        | APPLICATION PROBLEMS .....                       | 784 |

*Just as some women are said to be handsome  
though without adornment,  
so this subtle manner of speech,  
though lacking in artificial graces, delights us.*

Cicero (106–43 BC)

Models of randomly varying phenomena, even in idealized form, have been undeniably useful in providing solutions to many important practical problems — problems involving free-standing random variables from identifiable populations, as well as regression modeling for relating one variable to another. After making a case in the last chapter for taking more seriously the (oft-neglected) exercise of validating these models, and then presenting the means for carrying out these validation exercises, we are now faced with a very important question: what happens when the assumptions that are sup-

posed to make our data analysis lives easier are invalid? In particular, what happens when real life does not cooperate with the Gaussian distributional assumptions required for carrying out the *t*-tests and the *F*-tests, and other similar tests on which important statistical decisions rest?

Many tiptoe nervously around such issues, in the hope that the repercussions of invalid assumptions will be minimal; some stubbornly refuse to believe that violating any of these assumptions can really have any meaningful impact on their analyses; still others naïvely ignore such issues, primarily out of a genuine lack of awareness of the distributional requirements at the foundation of these analysis techniques. But none of this is an acceptable option for the well-trained engineer or scientist.

The objective in this chapter is to present some viable alternatives to consider when distributional assumptions are invalid. These techniques, which make little or no demand on the specific distributional structure of the population from whence the data came, are sometimes known as “distribution-free” methods. And precisely because they do not involve population parameters, in contrast to the distribution-based techniques, they are also known as nonparametric methods. Inasmuch as entire textbooks have been written on the subject of nonparametric statistics — complete treatises on statistical analysis without the support (and, some would say, encumbrance) of *hard* probability distribution models — the discussion here will necessarily be limited to only the few most commonly used techniques. And to put the techniques in proper context, we will compare and contrast these nonparametric alternatives with the corresponding parametric methods, where possible.

---

## 18.1 Introduction

There are at least two broad classes of problems for which the classical hypothesis tests discussed in Chapter 15 are unsuitable:

1. When the underlying distributional assumptions (especially the Gaussian assumptions) are seriously violated;
2. When the data in question is ordinal only, not measured on a quantitative scale in which the distance between succeeding entities is uniform or even meaningful (see Chapter 12).

In each of these cases, even in the absence of any knowledge of the mathematical characteristics of the underlying distributions, the sample data can always be rank ordered by magnitude. The data ranks can then be used to analyze such data with the little or no assumptions about the probability distributions of the populations.

**TABLE 18.1:** A professor's teaching evaluation scores organized by student type

| Graduate Students | Undergraduate Students |
|-------------------|------------------------|
| 3                 | 4                      |
| 4                 | 3                      |
| 4                 | 4                      |
| 2                 | 2                      |
| 3                 | 3                      |
| 4                 | 5                      |
| 4                 | 3                      |
| 5                 | 3                      |
| 2                 | 4                      |
| 4                 | 2                      |
| 4                 | 2                      |
| 4                 | 3                      |
|                   | 4                      |
|                   | 3                      |
|                   | 4                      |

Such nonparametric (or distribution-free) techniques have the obvious advantage that they are versatile: they can be used in a wide variety of cases, even when distributional assumptions are valid. By the same token, they are also quite robust (there are fewer or no assumptions to violate). However, for the very same reasons, they are not always the most efficient. For the same sample size, the power (the probability of correctly rejecting a false  $H_0$ ) is always higher for the parametric tests discussed in Chapter 15 *when the assumptions are valid*, compared to the power of a corresponding nonparametric test. Thus, if distributional assumptions are reasonably valid, parametric methods are preferred; when the assumptions are seriously in doubt, nonparametric methods provide a viable (perhaps even the only) alternative.

Finally, consider the case where a professor who taught an intermediate level statistics course to a class that included both undergraduate and graduate students, is evaluated on a scale from 1 to 5, where 5 is the highest rating. The evaluation scores from a class of 12 graduate and 15 undergraduate students is shown in Table 18.1. If the desire is to test whether or not the professor received more favorable ratings from graduate students, observe that this data set is ordinal only; it is not the usual quantitative data that is amenable to the usual parametric methods. But this ordinal data can be ranked since, regardless of the “distance” between each assigned number, we know that 5 is “better” than 4, which is in turn “better” than 3, etc. The question of whether graduate students rated the professor higher than under-

graduates can only be answered therefore by the nonparametric techniques we will discuss in this chapter.

The discussion to follow will focus on the following techniques:

- For single populations, the *one-sample sign test* and the *one-sample Wilcoxon signed rank test*;
- For two populations, the *Mann-Whitney-Wilcoxon (MWW) test*; and
- For probability model validation, the *Kolmogorov-Smirnov* and *Anderson-Darling* tests

Nonparametric tests for comparing more than two populations will be discussed at the appropriate places in the next chapter. Our presentation here will focus on the underlying principles, with some simple illustrations of the mechanics involved in the computations; much of the computational details will be left to, and illustrated with, computer programs that have facilitated the modern application of these techniques.

## 18.2 Single Population

### 18.2.1 One-Sample Sign Test

The one-sample sign test is a test of hypotheses about the median,  $\eta$ , of a continuous distribution. Recall that the median is defined as that value for which  $P(X \leq \eta) = P(X \geq \eta) = 0.5$ . The null hypothesis in this case is:

$$H_0 : \eta = \eta_0 \quad (18.1)$$

to be tested against the usual array of alternative:

$$\begin{aligned} H_{a_L} &: \eta < \eta_0 \\ H_{a_R} &: \eta > \eta_0 \\ H_{a_2} &: \eta \neq \eta_0 \end{aligned} \quad (18.2)$$

### Basic Test Characteristics

Given a random sample,  $X_1, X_2, \dots, X_n$ , as with other tests, first we need an appropriate test statistic to be used in determining rejection regions from such sample data. In this regard, consider the sample deviation from the postulated median,  $D_{mi}$ , defined by:

$$D_{mi} = X_i - \eta_0 \quad (18.3)$$

Furthermore, suppose that we are concerned for the moment *only* with the *sign* of this quantity, the magnitude being of no interest for now: i.e., all we care about is whether  $X_i$  shows a positive deviation from the postulated median (when  $X_i > \eta_0$ ) or it shows a negative deviation (when  $X_i < \eta_0$ ). Then,

$$D_{m_i} = \begin{cases} + & X_i \geq \eta \\ - & X_i \leq \eta \end{cases} \quad (18.4)$$

(Note that the requirement that  $X$  must be a continuous random variable, rules out, in principle — but not necessarily in practice — the potential for a “tie” where  $X_i$  exactly matches the value for the postulated median, since the probability of this event occurring is theoretically zero. However, if by chance  $X_i - \eta_0 = 0$ , this data is simply left out of the analysis.)

Observe that as defined in Eq (18.4), there are only two possible outcomes for the quantity,  $D_{m_i}$ , making it a classic Bernoulli random variable. Now, if  $H_0$  is true, the sequence of  $D_{m_i}$  observations should contain about an equal number of + and – entries. If  $T^+$  is the total number of + signs (representing the total number of positive deviations from the median, arising from observations greater than the median), then this random variable has a binomial distribution with binomial probability of “success”  $p_B = 0.5$  if  $H_0$  is true. Observe therefore that  $T^+$  has all the properties of a useful test statistic.

The following are therefore the primary characteristics of this test:

1. The test statistic is  $T^+$ , the total number of plus signs;
2. Its sampling distribution is binomial; specifically, if  $H_0$  is true,  $T^+ \sim Bi(n, 0.5)$

For any specific experimental data, the observed total number of plus signs,  $t^+$ , can then be used to compute the rejection region or, alternatively, to test directly for significance by computing  $p$ -values as follows. For the one-sided lower-tailed alternative, i.e.,  $H_{a_L}$ ,

$$p = P(T^+ \leq t^+ | p_B = 0.5) \quad (18.5)$$

the probability of the observing a total of  $t^+$  plus signs or fewer, out of a sample of  $n$ , given equiprobable + or – outcomes. For the one-sided upper-tailed alternative, i.e.,  $H_{a_R}$ , the required  $p$ -value is obtained from

$$p = P(T^+ \geq t^+ | p_B = 0.5) \quad (18.6)$$

Finally, for the two-tailed test,  $H_{a_2}$ ,

$$p = \begin{cases} 2P(T^+ \leq t^+ | p_B = 0.5); & \text{for } t^+ < n/2 \\ 2P(T^+ \geq t^+ | p_B = 0.5); & \text{for } t^+ > n/2 \end{cases} \quad (18.7)$$

Let us illustrate this procedure with an example.

**Example 18.1: MEDIAN OF EXPONENTIAL DISTRIBUTION**

The data set  $S_1 = \{16, 1, 9, 34, 63, 44, 1, 63, 41, 29\}$ , first presented in Example 14.3 (and later in Example 17.1), is the waiting time (in days) until the occurrence of a recordable safety incident in a certain company's manufacturing site. This was postulated to be from an exponential distribution. Use the one-sample sign test to test the null hypothesis that the median,  $\eta = 21$ , versus the two-sided alternative that it is not.

**Solution:**

From the given data, and the postulated median, we easily generate the following table:

| Time Data<br>( $S_1$ ) | Deviation<br>from $\eta_0$ | Sign<br>$D_{m_i}$ |
|------------------------|----------------------------|-------------------|
| 16                     | -5                         | -                 |
| 1                      | -20                        | -                 |
| 9                      | -12                        | -                 |
| 34                     | 13                         | +                 |
| 63                     | 42                         | +                 |
| 44                     | 23                         | +                 |
| 1                      | -20                        | -                 |
| 63                     | 42                         | +                 |
| 41                     | 20                         | +                 |
| 29                     | 8                          | +                 |

This shows six total plus signs, so that  $t^+ = 6$ . Because of the two-sided alternative hypothesis, and since  $t^+ > 5$ , we need to compute  $P(x \geq 6)$  for a  $Bi(10, 0.5)$ . This is obtained as:

$$P(x \geq 6) = 1 - P(x \leq 5) = 1 - 0.623 = 0.377 \quad (18.8)$$

The  $p$ -value associated with this sign test is therefore  $2 \times 0.377 = 0.754$ . Therefore, there is no evidence to reject the null hypothesis that the median is 21. (The sample median is 31.50, obtained as the average of the 5<sup>th</sup> and 6<sup>th</sup> ranked sample data, 29 and 34.)

It is highly recommended, of course, to use statistical software packages for carrying out such tests. To use MINITAB, upon entering the data into a column labeled S1, the sequence Stat > Nonparametrics > 1-Sample Sign opens the usual dialog box where the test characteristics are entered. The MINITAB output is as follows:

**Sign Test for Median: S1**  
**Sign test of median = 21.00 versus not = 21.00**

| N  | Below | Equal | Above | P | Median |       |
|----|-------|-------|-------|---|--------|-------|
| S1 | 10    | 4     | 0     | 6 | 0.7539 | 31.50 |

For large samples, the sampling distribution for  $T^+$ ,  $Bi(n, 0.5)$ , tends

to  $N(\mu, \sigma^2)$ , with mean,  $\mu = np_B = 0.5n$ , and standard deviation,  $\sigma = \sqrt{np_B(1 - p_B)}$ . Thus, the test statistic

$$Z = \frac{T^+ - 0.5n}{0.5\sqrt{n}} \quad (18.9)$$

can be used to carry out the sign test for large samples (typically  $> 10 - 15$ ), exactly like the standard  $z$ -test.

### Comparison with Parametric Alternatives

The sign test is the nonparametric alternative to the one-sample  $z$ -test and the one-sample  $t$ -test. These parametric tests are for hypotheses concerning the means of normal populations; the sign test on the other hand is for the median of any general population with a continuous probability distribution. If data is susceptible to outliers or the distribution is long-tailed (i.e., skewed), the sign test is more appropriate; if the distribution is close to normal, the parametric  $t$ - and  $z$ -tests will perform better.

#### 18.2.2 One-Sample Wilcoxon Signed Rank Test

The one-sample Wilcoxon signed rank test is also a test of hypotheses about the median of continuous distributions. It goes one step further than the sign test by taking into account the magnitude of the deviation from the median,  $D_{m_i}$ , in addition to the sign. However, it is restricted to symmetric distributions and therefore not applicable to skewed distributions. (The recommended option for skewed distributions is the more general, but less powerful, sign test discussed above.)

### Basic Test Characteristics

The test concerns the median,  $\eta$ , of a continuous and symmetric distribution, with the null and alternative hypotheses as stated in Eqs (18.1) and (18.2). From the usual sample data, obtain, as before,  $D_{m_i} = X_i - \eta_0$ , the deviation from the median. The test calls first for ranking the absolute  $|D_{m_i}|$  in ascending order, and subsequently attaching to these ranks, the signs corresponding to the original  $D_{m_i}$  values, to obtain the “signed ranks.”

Let  $W^+$  be the sum of the ranks with positive signs, (i.e., where  $D_{m_i} > 0$ ); and  $W^-$ , the sum of the ranks when  $D_{m_i} < 0$ . (For obvious reasons,  $W^+$  is known as the “positive rank sum.”) These are the basic statistics used for the Wilcoxon signed rank test. Different variations of the test use different versions of these test statistics. For example, some use one or the other of  $W^+$  or  $W^-$ , some use  $\max(W^+, W^-)$  and others  $\min(W^+, W^-)$ . MINITAB and several other software packages base the test on  $W^+$  as will the discussion in this section.

The statistic,  $W^+$ , has some distinctive characteristics: for example, for a sample size of  $n$ , the largest value it can attain is  $(1 + 2 + \dots + n) = n(n + 1)/2$

when the entire sample exceeds the postulated value for the median,  $\eta_0$ . It can attain a minimum value of 0. Between these extremes,  $W^+$  takes other values that are easily determined via combinatorial computations. And now, if  $H_0$  is true, the sampling distribution for  $W^+$  can be computed numerically and used to determine significance levels. The computations involved in determining the sampling distribution of  $W^+$  under  $H_0$  are cumbersome analytically, but relatively easy to execute with a computer program; the test is thus best carried out with statistical software packages.

As usual, large sample approximations exist. In particular, it can be shown that the sampling distribution for  $W^+$  tends to  $N(\mu, \sigma^2)$  for large  $n$ , with  $\mu = n(n + 1)/4$  and  $\sigma^2 = n(n + 1)(2n + 1)/24$ . But it is not recommended to use the normal approximation because it is not sufficiently precise; besides, the computations involved in the exact test are trivial for computer programs.

We use the next example to illustrate the mechanics involved in this test before completing the test itself with MINITAB.

**Example 18.2: CHARACTERIZING SOME HOUSEKEEPING GENES**

In genomic studies, genes that are involved in basic functions needed to keep the cell alive are always “turned on” (i.e., they are constitutively expressed). Such genes are known colloquially as “housekeeping genes.” Because their gene expression status hardly changes, they are sometimes used to “calibrate” experimental systems for measuring changes in gene expression. Data on 10 such putative housekeeping genes has been selected from a larger set of microarray data and presented as  $\phi$ , the fractional intensity,  $I_{test}/(I_{test} + I_{ref})$ , where  $I_{test}$  is the measured fluorescence intensity under test conditions, and  $I_{ref}$ , the intensity under reference conditions. Within the limits of random variability, the values of  $\phi$  determined for housekeeping genes should come from a symmetric distribution scaled between 0 and 1. If these 10 genes are true housekeeping genes, the median of the data population for  $\phi$  should be 0.5. To illustrate the mechanics involved in using the one-sample Wilcoxon signed rank test to test this hypothesis against the alternative that the median is not 0.5, the following table is a summary of the raw data and the subsequent analysis required for carrying out this test.

| $\phi_i$ | $D_{m_i} = \phi_i - 0.5$ | $ D_{m_i} $ | Rank | Signed Rank |
|----------|--------------------------|-------------|------|-------------|
| 0.585978 | 0.085978                 | 0.085978    | 5    | 5           |
| 0.504057 | 0.004057                 | 0.004057    | 1    | 1           |
| 0.182831 | -0.317169                | 0.317169    | 10   | -10         |
| 0.426575 | -0.073425                | 0.073425    | 4    | -4          |
| 0.455191 | -0.044809                | 0.044809    | 3    | -3          |
| 0.804720 | 0.304720                 | 0.304720    | 9    | 9           |
| 0.741598 | 0.241598                 | 0.241598    | 8    | 8           |
| 0.332909 | -0.167091                | 0.167091    | 7    | -7          |
| 0.532131 | 0.032131                 | 0.032131    | 2    | 2           |
| 0.610620 | 0.110620                 | 0.110620    | 6    | 6           |

The first column is the raw fractional intensity data; the second column is the deviation from the median whose absolute value is shown in column 3, and ranked in column 4. The last column shows the signed rank. Note that in this case,  $w^+ = 31$ , the sum of all the ranks carrying the plus sign, i.e.,  $(5 + 1 + 9 + 8 + 2 + 6)$ .

When MINITAB is used to carry out this test, the required sequence is **Stat** > **Nonparametrics** > **1-Sample Wilcoxon**; the result is shown below:

| <b>Wilcoxon Signed Rank Test: Phi</b>                     |                |                |       |        |
|---|----------------|----------------|-------|--------|
| <u>Test of median = 0.5000 versus median not = 0.5000</u> |                |                |       |        |
|   | N for Wilcoxon | Estimate       |       |        |
|   | N              | Test Statistic | P     | Median |
| Phi   | 10             | 31.0           | 0.760 | 0.5186 |

The high  $p$ -value indicates that there is no evidence to support rejecting the null hypothesis. We therefore conclude that the median is likely to be 0.5 and that the selected genes appear to be true housekeeping genes.

As with the sign test, occasionally  $X_i$  exactly equals the postulated median,  $\eta_0$ , creating a “tie,” even though the theoretical probability that this will occur is zero. Under these circumstance, most software packages simply set such data aside, and the output will reflect this. For example, in MINITAB, the “N for Test” entry indicates how many of the original sample of size N actually survived to be used for the test. In the gene expression example above there were no ties, since  $\phi$  was computed to a large enough number of decimal places.

### Comparison with Parametric Alternatives

The symmetric distribution requirement for the one-sample Wilcoxon signed rank test might make it appear to be a direct nonparametric alternative to the one-sample  $z$ - or  $t$ -tests, but this is not quite true. For normally distributed populations, the Wilcoxon signed rank test is slightly less powerful than the  $t$ -test. In any event, there is no reason to abandon the classical parametric tests when the Gaussian assumption is valid. It is for other symmetric distributions, such as the uniform, or the symmetric beta that the Wilcoxon signed rank test is particularly useful, where the Gaussian assumption does not readily hold.

---

### 18.3 Two Populations

The general problem involves two separate and *mutually independent* populations, with respective unknown medians  $\eta_1$  and  $\eta_2$ . As with the parametric

tests, we are typically concerned with testing hypotheses about the difference between these two medians, i.e.,

$$\eta_1 - \eta_2 = \delta. \quad (18.10)$$

Depending on the nature of the data sets, we can identify two categories: (i) the special case of paired data; and (ii) the more general case of unpaired samples.

### 18.3.1 Two-Sample Paired Test

In this case, once the differences between the pairs,  $D_i = X_{1i} - X_{2i}$  have been obtained, this can be treated exactly like the one-sample case presented above. The hypothesis will be tested on a postulated value for the median of  $D_i$ , say  $\delta_0$ , which need not be zero. When  $\delta_0 = 0$ , the test reduces to a test of the equality of the two population medians. In any event, the one-sample tests discussed above, either the sign test, or the Wilcoxon signed rank test (if the distribution of the difference,  $D_i$  can be considered as reasonably symmetric) can now be applied. No additional considerations are needed.

Thus, the two-sample paired test is exactly the same as the one-sample test when it is applied to the paired differences.

### 18.3.2 Mann-Whitney-Wilcoxon Test

The Mann-Whitney-Wilcoxon (MWW) test (also known as the two-sample Wilcoxon rank sum test) is a test of hypothesis regarding the median of two independent populations with identical, continuous distributions that are possibly shifted apart. The test is also applicable to discrete distributions provided the scale is ordinal. The typical null hypothesis, written in general form, is

$$H_0 : \eta_1 - \eta_2 = \delta_0 \quad (18.11)$$

where  $\eta_i$  is the median for distribution  $i$ . This is tested against the usual triplet of alternatives. For tests of equality,  $\delta_0 = 0$ .

#### Basic Test Characteristics

The test uses a random sample of size  $n_1$  from population 1,  $X_{11}, X_{12}, \dots, X_{1n_1}$ , and another of size  $n_2$  from population 2,  $X_{21}, X_{22}, \dots, X_{2n_2}$ , where  $n_1$  need not equal  $n_2$ . First, the entire  $n_T = n_1 + n_2$  observations are combined and ranked in ascending order as though they were from the same population. (Identical observations are assigned equal ranks determined as the average of the individual assigned ranks had they been different).

Two related statistics can now be identified:  $W_1$ , the sum of the ranks in sample 1, and  $W_2$  the equivalent sum for sample 2. Again, we note that

$$W_1 + W_2 = \frac{n_T(n_T + 1)}{2} \quad (18.12)$$

the sum of the first  $n_T$  integers. And now, because this sum is fixed, observe that a small value for one (adjusted for the possibility that  $n_1 \neq n_2$ ) automatically implies a large value for the other, and hence a large difference between the two. Also note that the maximum attainable value for  $W_1$  is  $n_1(n_1 + 1)/2$ , and for  $W_2$ ,  $n_2(n_2 + 1)/2$ . The Mann-Whitney  $U$  test statistic used to determine significance levels, is defined as follows

$$\begin{aligned} U_1 &= W_1 - \frac{n_1(n_1 + 1)}{2} \\ U_2 &= W_2 - \frac{n_2(n_2 + 1)}{2} \\ U &= \min(U_1, U_2) \end{aligned} \quad (18.13)$$

The original Wilcoxon test statistic is slightly different but leads to equivalent results; it derives from Eq (18.12), and makes use of the larger of the two rank sums, i.e.,

$$W_1 = \frac{n_T(n_T + 1)}{2} - W_2 \quad (18.14)$$

if  $W_1 > W_2$  and reversed if  $W_1 < W_2$ .

The sampling distribution for  $U$  or for  $W_1$ , (or  $W_2$ ) are all somewhat like that for  $W^+$  above; they can be determined computationally for any given  $n_1$  and  $n_2$  and used to compute significance levels. As such, the entire test procedure is best carried out with the computer.

We now use the next example to illustrate first the mechanics involved in carrying out this test, and then complete the test with MINITAB.

#### **Example 18.3: TESTING FOR SAFETY PERFORMANCE IMPROVEMENT**

Revisit the problem first presented in Example 14.3 regarding a company's attempt to improve its safety performance. Specifically, we recall that to improve its safety record, the company began tracking the time in between recordable safety incidents. During the first year of the program, the waiting time (in days) until the occurrence of a recordable safety incident was obtained as

$$S_1 = \{16, 1, 9, 34, 63, 44, 1, 63, 41, 29\}$$

The data record for the second year of the program is

$$S_2 = \{35, 26, 16, 23, 54, 13, 100, 1, 30, 31\}$$

On the basis of this data, which shows a mean time to incident as 30.1 days for the first year and 32.9 days for the second year, the company's safety coordinator is preparing to make the argument to upper management that there has been a noticeable improvement in the company's safety record from Year 1 to Year 2. Is there evidence to support this claim?

**Solution:**

First, from phenomenological reasonings, we expect the data to follow the exponential distribution. This was confirmed as reasonable in Chapter 17. When this fact is combined with a sample size that is relatively small, this becomes a clear case where the *t*-test is inapplicable. We could use distribution-appropriate interval estimation procedures to answer this question (as we did in Example 14.14 of Chapter 14). But such procedures are too specialized and involve custom computations.

Problems of this type are ideal for the Mann-Whitney-Wilcoxon test. A table of the data and a summary of the subsequent analysis required for this test is shown here.

|    | Year 1, $S_1$ | Rank | Year 2, $S_2$ | Rank |
|----|---------------|------|---------------|------|
| 16 | 6.5           | 35   | 14.0          |      |
| 1  | 2.0           | 26   | 9.0           |      |
| 9  | 4.0           | 16   | 6.5           |      |
| 34 | 13.0          | 23   | 8.0           |      |
| 63 | 18.5          | 54   | 17.0          |      |
| 44 | 16.0          | 13   | 5.0           |      |
| 1  | 2.0           | 100  | 20.0          |      |
| 63 | 18.5          | 1    | 2.0           |      |
| 41 | 15.0          | 30   | 11.0          |      |
| 29 | 10.0          | 31   | 12.0          |      |

First, note how the complete set of 20 total observations have been combined and ranked. The lowest entry, 1, occurs in 3 places — twice in sample 1, and once in sample 2. They are each assigned the rank of 2 which is the average of 1+2+3. The other ties, two entries of 16, and two of 63 are given respective ranks of 6.5 and 18.5.

Next, the sum of ranks for sample 1 is obtained as  $w_1 = 105.5$ ; the corresponding sum for sample 2 is  $w_2 = 104.5$ . Note that these sum up to 210, which is exactly  $n_T(n_T + 1)/2$  with  $n_T = 20$ ; the larger  $w_1$  is used to determine significance level. But even before the formal test is carried out, note the closeness of these two rank sums (the sample sizes are the same for each data set). This already gives us a clue that we are unlikely to find evidence of significant differences between the two medians.

To carry out the MWW test for this last example using MINITAB, the required sequence is **Stat > Nonparametrics > Mann-Whitney**; the ensuing results are shown below.

**Mann-Whitney Test and CI: S1, S2**

|    | N  | Median |
|----|----|--------|
| S1 | 10 | 31.5   |
| S2 | 10 | 28.0   |

Point estimate for ETA1-ETA2 is -0.00

95.5 Percent CI for ETA1-ETA2 is (-25.01,24.99)

W = 105.5

Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 1.0000

The test is significant at 1.0000 (adjusted for ties)

From here, we see that the  $p$ -value is 1.000, implying that there is no evidence to reject the null hypothesis,  $H_0 : \eta_1 = \eta_2$ . Notice that the W value reported for the test is the same as  $w_1$  obtained above, the larger of the two rank sums.

The next example illustrates the MWW test for discrete ordinal data. The primary point to watch out for in such cases is that because the data is discrete, there is a much higher chance that there will be ties.

**Example 18.4: MWW TEST FOR DISCRETE ORDINAL DATA**

From the data in Table 18.1, use MINITAB to test the hypothesis that the professor in question received equal ratings from both undergraduates and graduate students.

**Solution:**

The results of the MINITAB analysis is shown below:

**Mann-Whitney Test and CI: Grad, Undergrad**

|           | N  | Median |
|-----------|----|--------|
| Grad      | 12 | 4.000  |
| Undergrad | 15 | 3.000  |

Point estimate for ETA1-ETA2 is -0.000

95.2 Percent CI for ETA1-ETA2 is (0.000,1.000)

W = 188.0

Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.3413

The test is significant at 0.3106 (adjusted for ties)

Even though the median rating is 4.0 for the graduate students, and 3.0 for the undergraduates, the result, with  $p = 0.34$  ( $p = 0.32$  when adjusted for the ties), indicates that there is no evidence of a difference between the ratings given by the undergraduate students and those by graduate students.

### Comparison with Parametric Alternatives

The MWW test is generally considered as a direct nonparametric alternative to the two-sample  $t$ -test. When the populations are normal, the former test is somewhat less powerful than the latter with pooled sample variance; in most other cases, the MWW test is quite appreciably more powerful. However, if the two populations being compared are different (in shape and/or have different standard deviations, etc) the parametric two-sample  $t$ -test *without* pooling variances may be the better option.

## 18.4 Probability Model Validation

When the hypothesis test to be performed involves not just the mean, median or variance of a distribution but the entire distribution itself, the truly distribution-free approach is the Kolmogorov-Smirnov (K-S) test. The critical values of the test statistic (that determine the rejection region) are entirely independent of the specific distribution being tested. This makes the K-S test truly nonparametric; it also makes it less sensitive, especially to discrepancies between the observed data and the tail area characteristics of many distributions. The Anderson-Darling (A-D) test, a modified version of the K-S test designed to improve the K-S test sensitivity, achieves its improved sensitivity by making explicit use of the specific distribution to be tested in computing the critical values. This introduces the primary disadvantage that critical values must be custom-computed for each distribution. But what is lost in generality is more than made up for in overall improved performance.

We now review, briefly, the key characteristics of these two tests.

### 18.4.1 The Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov (K-S) test is a test concerning the distribution of a population from which a sample was obtained. It is based on the empirical cumulative distribution function determined from the sample data and applies only to continuous distributions.

#### Basic Test Characteristics

Let  $X_1, X_2, \dots, X_n$ , be a random sample drawn from a population with a postulated pdf  $f(x)$ , from which the corresponding cdf,  $F(x)$ , follows. If the random sample is ordered from the smallest to the largest as follows:  $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ , then, by definition, the empirical (data-based) probability that  $X \leq X_{(i)}$  is the ratio of the number of observations less than  $X_{(i)}$  (say  $n(i)$ ) to the total number of observations,  $n$ . Thus, the empirical cumulative distribution function is defined as:

$$F_E(i) = \frac{n(i)}{n} \quad (18.15)$$

The null and alternative hypotheses are:

- $H_0$  : The sample data follow specified distribution
- $H_a$  : The sample data do not follow specified distribution

Let  $F(x_{(i)})$  represent the theoretical cumulative probability  $P(X \leq x_{(i)})$  computed for a specific observation,  $x_{(i)}$ , using the theoretical cdf postulated for the population from which the data is purported to have come. The K-S test

is based on the difference between this theoretical cdf and the empirical cdf as in Eq (18.15). The test statistic is:

$$D = \max_{1 \leq i \leq n} \left\{ \left[ F(X_{(i)}) - \frac{i-1}{n} \right], \left[ \frac{1}{n} - F(X_{(i)}) \right] \right\} \quad (18.16)$$

As with other hypothesis tests, the null hypothesis is rejected at the significant level of  $\alpha$  if  $D > D_\alpha$ , where the critical value  $D_\alpha$  is typically obtained from computations easily carried out in software packages.

### Key Features

The primary distinguishing feature of the K-S test is that the sampling distribution of its test statistic,  $D$ , is entirely independent of the postulated distribution that is being tested. This makes the test truly distribution-free. It is a nonparametric alternative to the chi-squared goodness-of-fit test discussed in Chapter 17. Unlike that test which requires large sample sizes for the  $\chi^2$  distribution approximation for the test statistic to be valid, the K-S test is an exact test.

A key primary limitation is its restriction to continuous distributions that must be completely specified, i.e., the distribution parameters cannot be estimated from sample data. Also, because it is based on a single point with the “worst distance” between theoretical and empirical distributions, the test is prone to ignoring less prominent but still important mismatches at the tails, in favor of more influential mismatches at the center of the distribution. The K-S test is therefore more likely to have less power than a test that employs a more broadly based test statistic that is evenly distributed over the entire variable space. This is the *raison d'être* for the Anderson-Darling test which has all but supplanted the K-S test in many applications.

#### 18.4.2 The Anderson-Darling Test

The Anderson-Darling test<sup>1</sup> is a modification of the K-S test that (a) uses the entire cumulative distribution (not just the single worst point of departure from the empirical cdf), and hence, (b) gives more weight to the distribution tails than is possible with the K-S test.

The test statistic is:

$$A^2 = -n - \sum_{i=1}^n \frac{2i-1}{n} \{ \ln F(X_{(i)}) + \ln [1 - F(X_{(n+1-i)})] \}. \quad (18.17)$$

However, the sampling distribution of  $A^2$  depends on the postulated distribution function; critical values therefore must be computed individually for each distribution under consideration. Nevertheless, these critical values are

---

<sup>1</sup>Anderson, T. W.; Darling, D. A. (1952). “Asymptotic theory of certain ‘goodness-of-fit’ criteria based on stochastic processes”. *Annals of Mathematical Statistics* 23: 193212

available for many important distributions, including (but not limited to) the normal, lognormal, exponential, and Weibull distributions. The test is usually carried out using statistical software packages.

### Key Features

The primary distinguishing feature of the A-D test is that it is more sensitive than the K-S test, but for this advantage, the critical values for  $A^2$  must be custom-calculated for each distribution. The test is applicable even with small sample sizes,  $n < 20$ . It is therefore generally considered a better alternative to the chi-square and K-S goodness-of-fit tests.

## 18.5 A Comprehensive Illustrative Example

In experimental and theoretical neurobiology, “action potentials”—the spike trains generated by neurons—are used extensively to study nerve-cell activity. Because these spike trains and the dynamic processes that cause them are random, probability and statistics have been used to great benefit for such studies. For example, it is known that the distribution of interspike intervals (ISI)—the elapsed time between the appearance of two consecutive spikes in the spike train—can reveal something about synaptic mechanism<sup>2</sup>.

Table 18.2 contains 100 ISI measurements (in milliseconds) for the pyramidal tract cell of a monkey when awake (PT-W), and when asleep (PT-S), extracted from a more comprehensive set involving a few other cell types. The objective is to determine whether the activity of the pyramidal tract cell of a monkey is the same whether asleep or awake.

### 18.5.1 Probability Model Postulate and Validation

First, from phenomenological considerations, these interspike intervals represent time to the occurrence of several poisson events, suggesting that they might follow the gamma  $\gamma(\alpha, \beta)$  distribution, where  $\alpha$  will represent the effective number of poisson events leading to each observed action potential, and  $\beta$ , the mean rate of occurrence of the poisson events. A histogram of both data sets, with superimposed gamma distribution fits, is shown in Fig 18.1. Visually, the fits appear quite good, but we can make more quantitative statements by carrying out formal probability model validation.

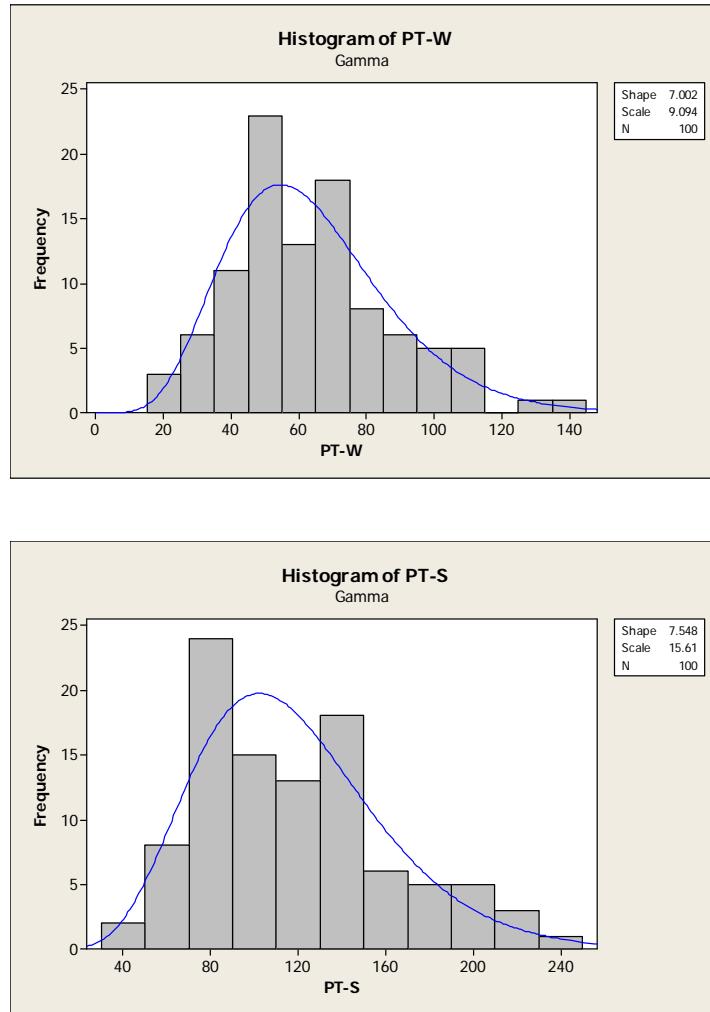
Because the postulated gamma model is continuous, we choose the probability plot approach and the Anderson-Darling test. The null hypothesis is that both sets of ISI data follow the gamma distribution; the alternative is

---

<sup>2</sup>Braitenberg, (1965): “What can be learned from spike interval histograms about synaptic mechanism?” *J. Theor. Biol.* **8**, 419–425

**TABLE 18.2:** Interspike intervals data for the pyramidal tract cell of a monkey when awake, (PT-W), and when asleep (PT-S)

| PT-W ISI (milliseconds) |     |     |     | PT-S ISI (milliseconds) |     |     |     |
|-------------------------|-----|-----|-----|-------------------------|-----|-----|-----|
| 25                      | 28  | 53  | 108 | 74                      | 135 | 132 | 94  |
| 49                      | 52  | 94  | 65  | 221                     | 137 | 125 | 146 |
| 56                      | 59  | 107 | 65  | 228                     | 80  | 173 | 179 |
| 52                      | 35  | 23  | 79  | 71                      | 116 | 44  | 195 |
| 48                      | 41  | 100 | 46  | 145                     | 119 | 99  | 156 |
| 113                     | 81  | 52  | 74  | 73                      | 55  | 150 | 192 |
| 50                      | 27  | 70  | 35  | 94                      | 88  | 236 | 143 |
| 76                      | 52  | 33  | 55  | 80                      | 143 | 157 | 78  |
| 68                      | 69  | 40  | 51  | 78                      | 121 | 49  | 199 |
| 108                     | 65  | 47  | 50  | 79                      | 130 | 162 | 98  |
| 45                      | 34  | 97  | 75  | 132                     | 68  | 139 | 138 |
| 73                      | 97  | 16  | 47  | 73                      | 97  | 152 | 187 |
| 105                     | 66  | 74  | 79  | 129                     | 66  | 202 | 63  |
| 60                      | 93  | 61  | 63  | 119                     | 111 | 151 | 86  |
| 70                      | 57  | 74  | 95  | 79                      | 85  | 56  | 104 |
| 64                      | 31  | 85  | 37  | 89                      | 136 | 105 | 116 |
| 39                      | 103 | 71  | 61  | 66                      | 105 | 60  | 133 |
| 92                      | 39  | 59  | 60  | 209                     | 175 | 96  | 137 |
| 49                      | 51  | 75  | 48  | 84                      | 157 | 81  | 89  |
| 94                      | 73  | 125 | 44  | 86                      | 133 | 178 | 116 |
| 71                      | 47  | 69  | 71  | 110                     | 103 | 73  | 85  |
| 95                      | 60  | 77  | 49  | 122                     | 89  | 145 | 90  |
| 54                      | 43  | 140 | 51  | 119                     | 94  | 98  | 222 |
| 54                      | 42  | 100 | 64  | 119                     | 141 | 102 | 81  |
| 56                      | 33  | 41  | 71  | 91                      | 95  | 75  | 63  |



**FIGURE 18.1:** Histograms of interspike intervals data with Gamma model fit for the pyramidal tract cell of a monkey. Top panel: when awake (PT-W); Bottom Panel: when asleep (PT-S). Note the *similarities* in the estimated values for  $\alpha$ —the shape parameter—for both sets of data, and the *difference* between the estimates for  $\beta$ , the scale parameters.

that they do not. The results of this test (carried out in MINITAB) are shown in Fig 18.2. The  $p$ -value for the A-D test in each case is higher than the typical significance level of 0.05, leading us to conclude that it is reasonable to consider the data as coming from gamma-distributed populations. The probability plots also show, for both cases, the entire data sets falling entirely within the 95% confidence intervals of the theoretical model line fits. The implication therefore is that we *cannot* use the two sample  $t$ -test for this problem, since the Gaussian assumption is invalid for confirmed gamma-distributed data. An examination of the histograms in fact shows two skewed distributions that have essentially the same shape, but the one for PT-S appears shifted to the right. The recommendation therefore is to use the Mann-Whitney-Wilcoxon test.

### 18.5.2 Mann-Whitney-Wilcoxon Test

Putting the data into two columns PT-W and PT-S in a MINITAB worksheet, and carrying out the test as illustrated earlier, yields the following results:

#### Mann-Whitney Test and CI: PT-W, PT-S

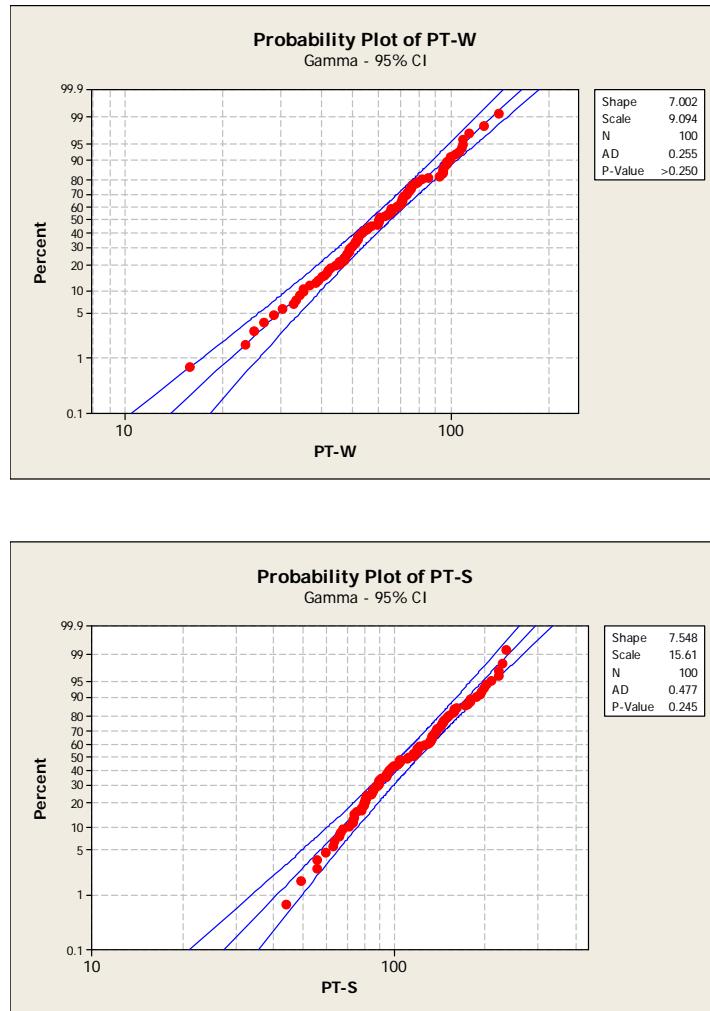
|      | N   | Median |
|------|-----|--------|
| PT-W | 100 | 60.01  |
| PT-S | 100 | 110.56 |

```
Point estimate for ETA1-ETA2 is -48.32
95.0 Percent CI for ETA1-ETA2 is (-59.24,-39.05)
W = 6304.0
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.0000
```

Since the significance level for the test (the  $p$ -value) is zero to four decimal places, the conclusion is that there is in fact a significant difference between the activities of these neurons.

A two-sample  $t$ -test may have led us to the same conclusion, but this would have been due more to the fortunate circumstance that the differences between the two activities are quite pronounced.

It is left as an exercise to the reader to provide some interpretations for what the estimated gamma model parameters might mean in terms of the phenomena underlying the generation of action potentials in these pyramidal tract neurons of the monkey.



**FIGURE 18.2:** Probability plot of interspike intervals data with postulated Gamma model and Anderson-Darling test for the pyramidal tract cell of a monkey. Top panel: when awake (PT-W); Bottom panel: when asleep (PT-S). The *p*-values for the A-D tests indicate no evidence to reject the null hypothesis

## 18.6 Summary and Conclusions

We have presented in this chapter some of the important nonparametric alternatives to consider when the assumptions underlying standard (parametric) hypothesis tests are not valid. These nonparametric techniques are applicable generally because they impose few, if any, restrictions on the data. They are known as “distribution free” methods because they do not rely on any specific distributional characterization for the underlying population. As a result, the primary variable of nonparametric statistical analysis is the *rank sum*, and for good reason: regardless of the underlying population from which they arose, all data—even qualitative data (so long as it is ordinal)—can be ranked, and the appropriate sums of such ranks provide valuable information about how the data set is distributed, without having to assume a specific functional form for the population’s distribution.

As summarized in Table 18.3, we have focussed specifically on:

- *The sign test*: for comparing a single continuous population median,  $\eta$ , to a postulated value,  $\eta_0$ . A nonparametric alternative to the one sample  $z$ - or  $t$ -test, it is based on the sign of the deviation of sample data from the postulated median. The test statistic,  $T^+$ —the total number of plus signs—is a binomial distributed  $Bi(n, 0.5)$  random variable if the null hypothesis is true.
- *The Wilcoxon signed rank test*: a more powerful version of the sign test because it takes both magnitude *and* sign of the deviation from the postulated median into consideration, but it is restricted to symmetric distributions. The test statistic,  $W^+$ , is the positive rank sum; its sampling distribution is best obtained numerically.
- *The Mann-Whitney-Wilcoxon (MWW) test*: for comparing the median of two independent populations with identical (but potentially shifted apart) continuous distributions; it is the nonparametric equivalent of the two-sample  $t$ -test. The test statistic,  $U$ , is based on  $W_1$  and  $W_2$ , the sums of the ranks in each sample.

The Kolmogorov-Smirnov (K-S) test and the Anderson-Darling (A-D) tests were also presented briefly for probability model validation, the latter being an improved version of the former. There are other nonparametric methods, of course, especially methods for nonparametric regression (e.g., Spearman’s rank correlation), but space limitations prevent us from discussing them all. The interested reader is referred, for example, to Gibbons and Chakraborti

(2003)<sup>3</sup>. The Kruskal-Wallis test, a nonparametric analog to the  $F$ -test, will be mentioned briefly at the appropriate place in Chapter 19.

It should be kept in mind, finally, that what makes the nonparametric tests versatile can also be a weakness. When underlying distributional information is available—especially when the Gaussian assumption is valid—it can be shown that nonparametric tests are not nearly as powerful: in order to draw conclusions with the same degree of confidence, larger samples will be required. Thus, whenever populations are reasonably normal, parametric tests are preferred. Some of the end-of-chapter exercises and problems illustrate this point.

This concludes the discussion of hypothesis testing begun in Chapter 15; and, in conjunction with the discussion of estimation (Chapters 13, 14, and 16) and of model validation (here and in Chapter 17), our treatment of what is generally considered as *statistical inference* is complete. However, there remains one more important issue: how to collect data such that the sample upon which statistical inference is to be based is as informative as possible. This is the focus of Chapter 19.

---

## REVIEW QUESTIONS

1. What is the objective of this chapter?
2. Why are the techniques discussed in this chapter known as “distribution free” methods?
3. What are the two broad classes of problems for which classical hypothesis tests of Chapter 15 are not applicable?
4. What are some of the advantages and disadvantages of non-parametric techniques?
5. What is the one-sample sign test?
6. What does it mean that a “tie” has occurred in a sign test, and what is done under such circumstances?
7. What is the test statistic used for the one-sample sign test, and what is its sampling distribution?
8. What is the large sample limiting test statistic for the one-sample sign test?

---

<sup>3</sup>Gibbons, J. D. and S. Chakraborti, (2003). *Nonparametric Statistical Inference*, 4th Ed. CRC Press.

TABLE 18.3: Summary of Selected Nonparametric Tests and their Characteristics

| Population Parameter   | Test                             | Restrictions                       | Test Statistic  | Parametric Alternative                           |
|--|----------------------------------|------------------------------------|---|--|
| $\eta; (H_0 : \eta = \eta_0)$<br>Median                            | Sign test                        | None                               | $T^+$ , Total # of positive deviations from $\eta_0$                            | One-sample $z$ -test<br>One-sample $t$ -test     |
| $\eta; (H_0 : \eta = \eta_0)$<br>Median                            | Wilcoxon signed Rank (WSR) test  | Continuous symmetric distributions | $W^+$ , Positive rank sum   | None   |
| $\delta = \eta_1 - \eta_2; (H_0 : \delta = \delta_0)$<br>(Paired)  | Sign test or WSR test            | Same as above                      | Same as above   | Same as above                                    |
| $\delta = \eta_1 - \eta_2; (H_0 : \delta = \delta_0)$<br>(General) | Mann-Whitney-Wilcoxon (MWW) test | Independent Distributions          | $W_1$ , Rank sum 1<br>$W_2$ , Rank sum 2<br>$U = f(W_1, W_2)$<br>See Eq (18.13) | Rank sum 1<br>Rank sum 2<br>Two-sample $t$ -test |

- 9.** The one-sample sign test is the nonparametric alternative to what parametric test?
- 10.** Apart from being distribution-free, what is another important distinction between the one-sample sign test and its parametric counterparts?
- 11.** When is the one-sample sign test more appropriate than the parametric alternatives? When will the parametric alternatives perform better?
- 12.** What is the one-sample Wilcoxon signed rank test, and what differentiates it from the one-sample sign test?
- 13.** What is the restriction on the Wilcoxon signed rank test?
- 14.** What is a “signed rank” and how is it obtained?
- 15.** What are the various test statistics that are used for the Wilcoxon signed rank test? In particular, what is  $W^+$ , the test statistic used in MINITAB and in this chapter?
- 16.** For a sample of size  $n$ , what are the minimum and maximum attainable values for the test statistic,  $W^+$ ?
- 17.** What is the large sample approximation to the sampling distribution of  $W^+$ ? Why is this approximation not recommended?
- 18.** As a result of the symmetric distribution restriction, can the Wilcoxon signed rank test be considered as a direct nonparametric alternative to the  $z$ - or  $t$ -test?
- 19.** For what symmetric distributions is the Wilcoxon signed rank test particularly useful?
- 20.** Why are no additional considerations needed for nonparametric two-sample paired tests?
- 21.** What is the Mann-Whitney-Wilcoxon test?
- 22.** What test statistics are involved in the Mann-Whitney-Wilcoxon test?
- 23.** Why is it especially important to rely on the computer for the Mann-Whitney-Wilcoxon test?
- 24.** The Mann-Whitney-Wilcoxon test is the nonparametric alternative to which parametric test? Under what conditions will one be better than the other?
- 25.** What is the Kolmogorov-Smirnov test used for?
- 26.** On what is the Kolmogorov-Smirnov test based?

- 27.** What are the null and alternative hypotheses in the Kolmogorov-Smirnov test?
- 28.** What is the primary distinguishing feature of the sampling distribution of the Kolmogorov-Smirnov test statistic?
- 29.** The Kolmogorov-Smirnov test is a non-parametric alternative to which parametric test?
- 30.** What are two primary limitations of the Kolmogorov-Smirnov test?
- 31.** How is the Anderson-Darling test related to the Kolmogorov-Smirnov test?
- 32.** The Anderson-Darling test is generally considered to be a better alternative to which tests?

## EXERCISES

**18.1** In the table below,  $X_1$  is a random sample of 20 observations from an exponential population with parameter  $\beta = 1.44$ , so that the median,  $\eta = 1$ .  $X_2$  is the same data set plus a constant, 0.6, and random Gaussian noise with mean  $\mu = 0$  and standard deviation  $\sigma = 0.15$ .

| $X_1$   | $X_2$   | $X_1$   | $X_2$   |
|---------|---------|---------|---------|
| 1.26968 | 1.91282 | 1.52232 | 2.17989 |
| 0.28875 | 1.13591 | 1.45313 | 2.11117 |
| 0.07812 | 0.72515 | 0.65984 | 1.45181 |
| 0.45664 | 1.19141 | 1.60555 | 2.45986 |
| 0.68026 | 1.34322 | 0.08525 | 0.43390 |
| 2.64165 | 3.18219 | 0.03254 | 0.76736 |
| 0.21319 | 0.88740 | 0.75033 | 1.16390 |
| 2.11448 | 2.68491 | 1.34203 | 2.01198 |
| 1.43462 | 2.16498 | 1.25397 | 1.80569 |
| 2.29095 | 2.84725 | 3.16319 | 3.77947 |

- (i) Consider the “postulate” that the median for both data sets is  $\eta_0 = 1$  (which, of course, is not true). Generate a table of signs,  $D_{mi}$ , of deviations from this postulated median.
- (ii) For each data set, determine the test statistic,  $T^+$ . What percentage of the observations in each data set has plus signs? Informally, what does this indicate about the possibility that the true median in each case is as postulated?

**18.2** Refer to Exercise 18.1 and the supplied data.

- (i) For each data set,  $X_1$  and  $X_2$ , carry out a formal sign test of the hypothesis  $H_0 : \eta = 1$  against  $H_a : \eta \neq 1$ . Interpret your results.
- (ii) Now use only the first 10 observations (on the left) in each data set. Repeat (i) above. Interpret your results. Since we know that the two data sets are different, and that the median for  $X_2$  is higher by 0.6 (on average), comment on the effect of

sample size on this particular sign test.

**18.3** Refer to Exercise 18.1 and the supplied data.

(i) It is known that the distribution of the difference of two exponentially distributed random variables is the *symmetric* Laplace distribution (See Exercise 9.3). This suggests that the one-sample Wilcoxon signed rank test, which is not applicable to either  $X_1$  or  $X_2$ , being samples from non-symmetric distributions, is applicable to  $D = X_1 - X_2$ . Carry out a one-sample Wilcoxon signed rank test on  $D$  to test the null hypothesis  $H_0 : \eta_D = 0$ , where  $\eta_D$  is the median of  $D$ , versus the alternative  $H_a : \eta_D \neq 0$ . Interpret your result in light of what we know about how the two data sets were generated.

(ii) Carry out a Mann-Whitney-Wilcoxon (MWW) two-sample test directly on the two samples  $X_1$  and  $X_2$ . Discuss the difference between the results obtained here and the ones obtained in (i).

(iii) How are the results obtained in this exercise reminiscent of the difference between the parametric standard two-sample  $t$ -test and a paired  $t$ -test discussed in Chapter 15?

**18.4** In Exercise 17.4, the two data sets in the table below were presented as random samples from two independent lognormal distributions; specifically,  $X_{L_1} \sim \mathcal{L}(0, 0.25)$  and  $X_{L_2} \sim \mathcal{L}(0.25, 0.25)$ .

(i) From these postulated theoretical distribution characteristics, determine the median of each population,  $\eta_{01}$  for  $X_{L_1}$  and  $\eta_{02}$  for  $X_{L_2}$ . Carry out a one-sample sign test that the median of the sample data  $X_{L_1}$  is  $\eta_{01}$  versus the alternative that it is not. Also carry out a one-sample sign test, this time that the median of the sample data  $X_{L_2}$  is also  $\eta_{01}$  versus the alternative that it is not. Is this test able to detect that  $X_{L_2}$  does not have the same median as  $X_{L_1}$ ?

(ii) Carry out an appropriate log transformation of the data to obtain  $Y_1$  and  $Y_2$  respectively from  $X_{L_1}$  and  $X_{L_2}$ . Determine the *theoretical* means and variances for the postulated populations of the log-transformed data, respectively  $(\mu_1, \sigma_1^2)$  and  $(\mu_2, \sigma_2^2)$ . Carry out a one-sample  $z$ -test that the mean of  $Y_1$  is the just-determined  $\mu_1$  value, versus the alternative that it is not. Also carry out a second one-sample  $z$ -test, this time that the mean of  $Y_2$  is the same as the just-determined  $\mu_1$  value, versus the alternative that it is not.

(iii) Comment on any differences observed between the results of the sign test on the raw data and the  $z$ -test on the transformed data.

| $X_{L_1}$ | $X_{L_2}$ |
|-----------|-----------|
| 0.81693   | 1.61889   |
| 0.96201   | 1.15897   |
| 1.03327   | 1.17163   |
| 0.84046   | 1.09065   |
| 1.06731   | 1.27686   |
| 1.34118   | 0.91838   |
| 0.77619   | 1.45123   |
| 1.14027   | 1.47800   |
| 1.27021   | 2.16068   |
| 1.69466   | 1.46116   |

**18.5** Refer to Exercise 18.4. (i) Carry out a MWW test on the equality of the medi-

ans of  $X_{L_1}$  and  $X_{L_2}$ , versus the alternative that the medians are different. Interpret your result.

- (ii) Carry out a two-sample  $z$ -test concerning the equality of the means of the log transformed variables  $Y_1$  and  $Y_2$ , against the alternative that the means are different. Interpret your results.
- (iii) Comment on any differences observed between these two sets of results.

**18.6** The data in the table below are from two populations that may or may not be the same.

| $X_U$ | $Y_U$ |
|-------|-------|
| 0.65  | 1.01  |
| 2.01  | 1.75  |
| 1.80  | 1.27  |
| 1.13  | 2.48  |
| 1.74  | 2.91  |
| 1.36  | 2.38  |
| 1.55  | 2.79  |
| 1.05  | 1.94  |
| 1.55  |       |
| 1.63  |       |

- (i) Carry out a two-sample  $t$  test to compare the population means. What assumptions are required for this to be a valid test? At the  $\alpha = 0.05$  significance level, what does this result imply about the null hypothesis?
- (ii) Carry out a MWW test to determine if the two populations have the same medians or not. At the  $\alpha = 0.05$  significance level, what does this result imply about the null hypothesis?
- (iii) It turns out that the data were generated from two distinct uniform distributions, where the  $Y$  distribution is slightly different. Which test, the parametric or the nonparametric, is more effective in this case? Offer some reasons for the observed performance of one test versus the other.
- (iv) In light of what was specified about the two populations in (iii), and the  $p$ -values associated with each test, comment on the use of  $\alpha = 0.05$  as an *absolute* arbiter of significance.

**18.7** In an opinion survey on a particular political referendum, fifteen randomly samples individuals were asked to give their individual opinions using the following options:

1= Strongly agree; 2 = Agree; 3 = Indifferent; 4 = Disagree; 5 = Strongly disagree. The result is the data set,  $S_{15}$ .

$$S_{15} = \{1, 4, 5, 3, 2, 3, 3, 2, 2, 2, 1, 3, 3, 3, 3\}$$

Carry out a one-sample sign test to confirm or refute the allegation that the population from which the 15 respondents were sampled has a median of people that are indifferent to the referendum in question.

**18.8** The following is a set of residuals from a two-parameter regression model:

|       |       |      |       |       |       |       |      |      |       |
|-------|-------|------|-------|-------|-------|-------|------|------|-------|
| 0.97  | -0.12 | 0.27 | 0.43  | 0.17  | 0.10  | -0.58 | 0.96 | 0.04 | 0.54  |
| -0.68 | 0.04  | 0.33 | -0.04 | -0.34 | -0.85 | 0.12  | 0.14 | 0.49 | -0.09 |

(i) Carry out a K-S test of normality and compare it with an A-D test of normality. What are the associated  $p$ -values and what do these imply about the normality of these residuals?

(ii) It turns out that the residuals shown above had been “filtered” by removing four observations that appeared to be outliers:  $\{1.30, -1.75, 2.10, -1.55\}$ . Reinstate these residuals and repeat (i). Does this change any of the conclusions about the normality of the residuals?

**18.9** The following data is from a population that may or may not be normally distributed.

|       |       |      |       |      |       |      |       |       |      |
|-------|-------|------|-------|------|-------|------|-------|-------|------|
| 0.88  | 2.06  | 6.86 | 1.42  | 2.42 | -0.29 | 0.74 | -0.91 | -2.30 | 0.32 |
| -0.74 | -0.30 | 1.06 | -3.08 | 1.12 | 0.63  | 0.58 | -0.48 | -3.15 | 6.70 |

Carry out a K-S test of normality and compare the results with an A-D test of normality. At the  $\alpha = 0.05$  significance level, what do each of these tests imply about the normality of the data set? Examine the data set carefully and comment on which of these tests is more likely to lead to the correct decision.

## APPLICATION PROBLEMS

**18.10** In Problem 15.59, hypothesis tests were devised to ascertain whether or not out of three operators, “A,” “B,” and “C,” working in a toll manufacturing facility, operator “A” was deemed to be more safety conscious. The data below, showing the time in months between occurrences of safety violations for each operator, was to be used to test these hypotheses.

|   |      |      |      |      |      |      |      |      |      |      |
|---|------|------|------|------|------|------|------|------|------|------|
| A | 1.31 | 0.15 | 3.02 | 3.17 | 4.84 | 0.71 | 0.70 | 1.41 | 2.68 | 0.68 |
| B | 1.94 | 3.21 | 2.91 | 1.66 | 1.51 | 0.30 | 0.05 | 1.62 | 6.75 | 1.29 |
| C | 0.79 | 1.22 | 0.65 | 3.90 | 0.18 | 0.57 | 7.26 | 0.43 | 0.96 | 3.76 |

Unfortunately, the random variable in question is exponentially distributed; the sample size of 10 is considerably smaller than is required for a normal approximation to be valid for the sampling distribution of the sample mean; and therefore no standard hypothesis test could be used.

Convert the data to a form that will be appropriate for carrying out a one-sample Wilcoxon signed rank test to compare operator “A” to operator “B,” and a second one-sample Wilcoxon signed rank test to compare operator “A” to operator “C.” Interpret your results. Can operator “A” truly be said to be more safety conscious than either “B” or “C”? If yes, at what significance level?

**18.11** Refer to Problem 18.10.

(i) Carry out a two-sample  $t$ -test to compare the mean time between occurrences of safety violations for operator “A” to that for operator “B”; carry out a second

two-sample  $t$ -test comparing the means for operator “A” to that for operator “C”. Are these valid tests?

(ii) Carry out a MW<sub>W</sub> test to compare the safety performance of operator “A” to that for operator “B” and a second MW<sub>W</sub> test to compare operator “A” to operator “C”. Are these valid tests? Interpret your results and compare them to the results in (i). What conclusions can you reach about these operators and how safety conscious “B” and “C” are in comparison to “A”?

**18.12** The Philadelphia Eagles, like every other team in the National Football League (NFL) in the US, plays 8 games at home and 8 games away in each 16-game season. The table below shows the total number of points scored by this NFL team at home and away during the 2008/2009 season.

| Total Points Scored, 2008/2009 Season |    |    |    |    |    |    |    |    |  |
|---------------------------------------|----|----|----|----|----|----|----|----|--|
| Home                                  | 38 | 15 | 17 | 27 | 31 | 48 | 30 | 44 |  |
| Away                                  | 37 | 20 | 40 | 26 | 13 | 7  | 20 | 3  |  |

(i) Generate side-by-side box-plots of the two data sets and comment on what these plots shows about the potential difference between the number of points scored at home and away.

(ii) Use the two-sample  $t$ -test to compare the mean offensive productivity at home versus away. What assumptions are necessary for this to be a valid test? Are these assumptions reasonable? Interpret your result.

(iii) Next, carry out a MW<sub>W</sub> test. Interpret your result.

(iv) Allowing for the fact that there are only 8 observations for each category, discuss your personal opinion of what the data set implies about the difference in offensive output at home versus away, *vis à vis* the results of the formal tests.

**18.13** Refer to Problem 18.12. This time, the table below shows the point differential—points scored by the Philadelphia Eagles minus points scored by the opponent—at home and away, during the 2008/2009 season. Some consider this metric a better measure of ultimate team performance (obviously, a negative point differential corresponds to a loss, a positive differential a win, and a zero, a tie).

| Point Differential, 2008/2009 Season |    |    |    |    |    |     |    |    |
|--------------------------------------|----|----|----|----|----|-----|----|----|
| Home                                 | 35 | 9  | -6 | 13 | -5 | 28  | 20 | 38 |
| Away                                 | -4 | -4 | 14 | 19 | 0  | -29 | 6  | -7 |

(i) Generate side-by-side box-plots of the two data sets and comment on what this plot shows about the potential difference between the team’s performance at home and away.

(ii) Carry out a two-sample  $t$ -test to compare the team’s performance at home versus away. What assumptions are necessary for this to be a valid test? Are these assumptions reasonable? Interpret your results.

(iii) Next, carry out a MW<sub>W</sub> test. Interpret your result.

(iv) Again, allowing for the fact that there are only 8 observations in each case, discuss your personal opinion about the difference in team performance at home versus away *vis à vis* the results of the formal tests.

**18.14** The table below shows the result of a market survey involving 15 women and

15 men who were asked to taste a “name brand” diet Cola drink and compare the taste to that of a generic brand that is cheaper, but, as claimed by the manufacturer, whose taste is preferred by a majority of tasters. The options given to the participants are as follows: 1 = Strongly prefer generic cola; 2 = Prefer generic cola; 3 = Indifferent; 4 = Prefer name brand cola; 5 = Strongly prefer name brand cola.

Perform appropriate tests to validate the claims that (i) Men are mostly indifferent, showing no preference one way or another (which is a “positive” result for the generic cola manufacturer); and (ii) there is no difference between Women and Men in their preferences for the diet Cola brands. Is there evidence in the data to support one or both or none of these claims?

| Women | Men |
|-------|-----|
| 4     | 1   |
| 3     | 3   |
| 5     | 3   |
| 5     | 2   |
| 2     | 4   |
| 3     | 5   |
| 1     | 1   |
| 4     | 2   |
| 4     | 4   |
| 5     | 3   |
| 4     | 3   |
| 3     | 2   |
| 4     | 1   |
| 3     | 3   |
| 4     | 3   |

---

**18.15** Random samples of size 10 each are taken from large groups of trainees instructed by Method A and Method B, and each trainee’s score on an appropriate achievement test is shown below.

|          |    |    |    |    |    |    |    |    |    |    |
|----------|----|----|----|----|----|----|----|----|----|----|
| Method A | 71 | 75 | 65 | 69 | 73 | 66 | 68 | 71 | 74 | 68 |
| Method B | 72 | 77 | 84 | 78 | 69 | 70 | 77 | 73 | 65 | 75 |

In Example 15.6, the data sets were assumed to come from normal populations and a two-sample  $t$ -test was conducted to test the claim that Method B is more efficient. As an alternative to that parametric test, carry out a corresponding MWU test. Interpret your result and compare it to that in Example 15.6. Is there a difference in these results? Comment on which test you would consider more reliable and why.

**18.16** Tanaka-Yamawaki (2003)<sup>4</sup>, presented models of high-resolution financial time series which showed, among other things, that price fluctuations tend to follow the Cauchy distribution, not the Gaussian distribution as usually presumed. The following table shows a particular sequence of similar price fluctuations.

---

<sup>4</sup>Mieko Tanaka-Yamawaki, (2003). “Two-phase oscillatory patterns in a positive feedback agent model” *Physica A* 324, 380–387

|           |           |           |           |
|-----------|-----------|-----------|-----------|
| -0.003322 | -0.000637 | 0.003569  | -0.032565 |
| 0.000856  | -0.002606 | -0.003797 | -0.001522 |
| 0.010382  | 0.061254  | -0.261127 | -0.032485 |
| 0.011494  | 0.004949  | 0.005694  | 0.034964  |
| 0.012165  | -0.008889 | 0.023339  | -0.009220 |

- (i) Obtain a histogram and a box plot of this data. Interpret these plots *vis à vis* the usual normality assumption of data of this sort.
- (ii) Carry out a K-S test and also an A-D test of normality on this data. What can you conclude about the normality of this data set?
- (iii) The data entry, -0.261127, although real, is what most might refer to as an “outlier” which will then be removed. Remove this data point and repeat (i) and (ii). Comment on the influence, if any, of this point on your analysis.

**18.17** The number of accidents occurring per quarter (three months) at a DuPont company facility, over a 10-year period is shown in the table below<sup>5</sup>, partitioned into two periods: Period I for the first five-year period of the study; Period II, the second five-year period.

| Period I |   |    |    | Period II |   |   |   |
|----------|---|----|----|-----------|---|---|---|
| 5        | 5 | 10 | 8  | 3         | 4 | 2 | 0 |
| 4        | 5 | 7  | 3  | 1         | 3 | 2 | 2 |
| 2        | 8 | 6  | 9  | 7         | 7 | 1 | 4 |
| 5        | 6 | 5  | 10 | 1         | 2 | 2 | 1 |
| 6        | 3 | 3  | 10 | 4         | 4 | 4 | 4 |

The phenomenon in question is clearly Poisson, so that it is not exactly valid to consider the data as samples from a normal population (even though with a sample size of 20 each, the distribution of the sample mean may be approximately Gaussian).

Carry out a MWU test to determine whether by Period II, there has been a significant improvement in the number of accidents occurring at this facility. Interpret your results adequately. Strictly speaking, is this a valid application of the MWU test? Why or why not?

**18.18** A limousine company that operates in a large metropolitan city in the North East of the United States has traditionally used in its fleet only two brands of American-made luxury cars. (To protect the manufacturers, we shall refer to these simply as brands *A* and *B*.) For various reasons, the company wishes to consolidate its operations and use only one brand. The decision is to be based on which brand experienced the fewest number of breakdowns over the immediately preceding five-year period. Seven cars per brand were selected randomly from the fleet, and from their maintenance records over this period, the following information was gathered on the number of breakdowns experienced by each car. Carry out an appropriate analysis of the data and recommend, objectively, which car brand, *A* or *B*, the company should select.

<sup>5</sup>Lucas J. M., (1985). “Counted Data CUSUMs,” *Technometrics*, 27, 129–144

| Total # of Breakdowns |         |
|-----------------------|---------|
| Brand A               | Brand B |
| 11                    | 9       |
| 9                     | 12      |
| 12                    | 10      |
| 14                    | 8       |
| 13                    | 4       |
| 12                    | 10      |
| 11                    | 8       |

**18.19** In Problem 15.52 (see also Problems 1.13 and 14.42) the data in the following table was analyzed to test a hypothesis about the mean time-to-publication for papers sent to a particular leading chemical engineering research journal. The data shows the time (in months) from receipt to publication of 85 papers published in the January 2004 issue of this journal.

- (i) On phenomenological grounds, and from past experience, a gamma distribution has been postulated for the population from which the data was sampled. The population parameters are unavailable. Carry out an A-D test to validate this probability model postulate.
- (ii) In recognition of the fact that the underlying distribution is obviously skewed, the Editor-in-Chief has modified his hypothesis about the characteristic time-to-publication, and now proposes that the *median* time-to-publication is 8 months. Carry out an appropriate test to assess the validity of this statement against the alternative that the *median* time-to-publication is *not* 8 months. Interpret your results.
- (iii) Repeat the test in (ii), this time against the alternative that the *median* time-to-publication is *longer than* 8 months. Reconcile this result with the one in (ii).

|      |      |      |      |      |
|------|------|------|------|------|
| 19.2 | 15.1 | 9.6  | 4.2  | 5.4  |
| 9.0  | 5.3  | 12.9 | 4.2  | 15.2 |
| 17.2 | 12.0 | 17.3 | 7.8  | 8.0  |
| 8.2  | 3.0  | 6.0  | 9.5  | 11.7 |
| 4.5  | 18.5 | 24.3 | 3.9  | 17.2 |
| 13.5 | 5.8  | 21.3 | 8.7  | 4.0  |
| 20.7 | 6.8  | 19.3 | 5.9  | 3.8  |
| 7.9  | 14.5 | 2.5  | 5.3  | 7.4  |
| 19.5 | 3.3  | 9.1  | 1.8  | 5.3  |
| 8.8  | 11.1 | 8.1  | 10.1 | 10.6 |
| 18.7 | 16.4 | 9.8  | 10.0 | 15.2 |
| 7.4  | 7.3  | 15.4 | 18.7 | 11.5 |
| 9.7  | 7.4  | 15.7 | 5.6  | 5.9  |
| 13.7 | 7.3  | 8.2  | 3.3  | 20.1 |
| 8.1  | 5.2  | 8.8  | 7.3  | 12.2 |
| 8.4  | 10.2 | 7.2  | 11.3 | 12.0 |
| 10.8 | 3.1  | 12.8 | 2.9  | 8.8  |

**18.20** Many intrinsic characteristics of single crystals affect the intensities of X-rays diffracted from them in different directions. The statistical distribution of these intensity measurements therefore provides means by which to characterize these crystals. Unfortunately, because of their distinctive “heavy tails,” these distributions are not adequately described by traditional default Gaussian distributions, which

is why Cauchy distributions have been finding increasing application in crystallographic statistics. (See for example Mitra and Sabita, (1989)<sup>6</sup>, and (1992)<sup>7</sup>.)

The table below is extracted from a larger sample of X-ray intensity measurements (arbitrary units) for two crystals that are very similar in structure, one labeled *A* is natural, the other, *B*, synthetic. The synthetic crystal is being touted as a replacement for the natural one.

| X-ray intensities (AU) |         |
|------------------------|---------|
| $XRD_A$                | $XRD_B$ |
| 104.653                | 132.973 |
| 106.115                | 114.505 |
| 104.716                | 115.735 |
| 104.040                | 114.209 |
| 105.631                | 114.440 |
| 104.825                | 100.344 |
| 117.075                | 116.067 |
| 105.143                | 114.786 |
| 105.417                | 115.015 |
| 98.574                 | 99.537  |
| 105.327                | 115.025 |
| 104.877                | 115.120 |
| 105.637                | 107.612 |
| 105.305                | 116.595 |
| 104.291                | 114.828 |
| 100.873                | 114.643 |
| 106.760                | 113.945 |
| 105.594                | 113.974 |
| 105.600                | 113.898 |
| 105.211                | 125.926 |
| 105.559                | 114.952 |
| 105.583                | 117.101 |
| 105.357                | 113.825 |
| 104.530                | 117.748 |
| 101.097                | 116.669 |
| 105.381                | 114.547 |
| 104.528                | 113.829 |
| 105.699                | 115.264 |
| 105.291                | 116.897 |
| 105.460                | 113.026 |

- (i) Carry out and A-D normality tests on each data set and confirm that these distributions, even though apparently symmetric, are *not* Gaussian. Their heavy tails imply that the Cauchy distribution may be more appropriate.
- (ii) The signature characteristic of the Cauchy distribution, that none of its moments exists, has the serious implication for statistical inference that, unlike other “non-pathological” distributions for which the sampling distribution of the mean narrows with increasing sample size, the distribution of the mean of a Cauchy sample is

<sup>6</sup>Mitra, G.B. and D. Sabita (1989). “Cauchy Distribution, Intensity Statistics and Phases of Reflections from Crystal Planes.” *Acta Crystallographica A* 45, 314-319.

<sup>7</sup>Mitra, G.B. and D. Sabita (1992). “Cauchy distribution of X-ray intensities: a note on hypercentric probability distributions of normalized structure factors.” *Indian Journal of Physics* 66 A (3), 375-378.

precisely the same as the original mother distribution. As a result, none of the standard parametric tests can be used for samples from Cauchy (and Cauchy-like) distributions.

From the sample data provided here, carry out an appropriate nonparametric test to ascertain the validity of the proposition that the synthetic crystal "B" is the same as the natural crystal "A", strictly on the basis of the X-ray intensity measurement.

# **Chapter 19**

---

## ***Design of Experiments***

|        |  |     |
|--------|--|-----|
| 19.1   | Introductory Concepts .....                    | 792 |
| 19.1.1 | Experimental Studies and Design .....          | 793 |
| 19.1.2 | Phases of Efficient Experimental Studies ..... | 794 |
| 19.1.3 | Problem Definition and Terminology .....       | 795 |
| 19.2   | Analysis of Variance .....                     | 795 |
| 19.3   | Single Factor Experiments .....                | 797 |
| 19.3.1 | One-Way Classification .....                   | 797 |
|        | Postulated Model and Hypotheses .....          | 797 |
|        | Data Layout and Experimental Design .....      | 798 |
|        | Analysis .....                                 | 799 |
|        | Fixed, Random, and Mixed Effects .....         | 803 |
|        | Summary Characteristics .....                  | 803 |
| 19.3.2 | Kruskal-Wallis Nonparametric Test .....        | 805 |
| 19.3.3 | Two-Way Classification .....                   | 805 |
|        | The Randomized Complete Block Design .....     | 805 |
|        | Postulated Model and Hypotheses .....          | 806 |
|        | Data Layout and Experimental Design .....      | 806 |
|        | Analysis .....                                 | 807 |
| 19.3.4 | Other Extensions .....                         | 810 |
| 19.4   | Two-Factor Experiments .....                   | 811 |
|        | Postulated Model and Hypotheses .....          | 811 |
|        | Data Layout and Experimental Design .....      | 812 |
|        | Analysis .....                                 | 812 |
| 19.5   | General Multi-factor Experiments .....         | 812 |
| 19.6   | $2^k$ Factorial Experiments and Design .....   | 814 |
| 19.6.1 | Overview .....                                 | 814 |
|        | Notation and Terminology .....                 | 815 |
|        | Characteristics .....                          | 815 |
| 19.6.2 | Design and Analysis .....                      | 816 |
| 19.6.3 | Procedure .....                                | 817 |
|        | Sample Size Considerations .....               | 818 |
| 19.6.4 | Closing Remarks .....                          | 821 |
| 19.7   | Screening Designs: Fractional Factorial .....  | 821 |
| 19.7.1 | Rationale .....                                | 822 |
| 19.7.2 | Illustrating the Mechanics .....               | 822 |
| 19.7.3 | General characteristics .....                  | 823 |
|        | Notation and Alias Structure .....             | 823 |
|        | Design Resolution .....                        | 824 |
| 19.7.4 | Design and Analysis .....                      | 825 |
|        | Basic Principles .....                         | 825 |
|        | Projection and Folding .....                   | 825 |
| 19.7.5 | A Practical Illustrative Example .....         | 827 |
|        | Problem Statement .....                        | 827 |
|        | Design and Data Collection .....               | 827 |
|        | Analysis Part 1 .....                          | 828 |

|   |     |
|---|-----|
| Analysis Part II: Projection .....            | 831 |
| 19.8 Screening Designs: Plackett-Burman ..... | 832 |
| 19.8.1 Primary Characteristics .....          | 833 |
| 19.8.2 Design and Analysis .....              | 833 |
| 19.9 Response Surface Designs .....           | 834 |
| 19.9.1 Characteristics .....                  | 834 |
| 19.9.2 Response Surface Designs .....         | 835 |
| 19.9.3 Design and Analysis .....              | 836 |
| 19.10 Introduction to Optimal Designs .....   | 837 |
| 19.10.1 Background .....                      | 837 |
| 19.10.2 "Alphabetic" Optimal Designs .....    | 837 |
| 19.11 Summary and Conclusions .....           | 839 |
| REVIEW QUESTIONS .....                        | 840 |
| EXERCISES .....                               | 842 |
| APPLICATION PROBLEMS .....                    | 849 |

*We may have three main objects in the study of truth:  
first, to find it when we are seeking it;  
second, to demonstrate it after we have found it;  
third, to distinguish it from error by examining it.*

Blaise Pascal (1623–1662)

Every experimental investigation presupposes that the sought-after information is contained in the acquired data sets. Objectively, however, such presupposition is often unjustifiable. Without giving purposeful, deliberate and careful consideration to data collection, the information content of the acquired data set cannot be guaranteed. Because experimental data will always be finite samples drawn from the much larger population of interest, conclusions drawn about the population will be valid *only* if the sample is representative. And the sample will be representative only if it encapsulates relevant population information appropriately. These issues have serious consequences. If the sought-after information is not contained in the data, no analysis technique — no matter how sophisticated — will be able to extract it. Therefore, how data sets are obtained will affect not only the information they contain, but also our ability to extract and use this information.

This chapter focusses on how experiments can be designed to ensure that acquired data sets are as informative as possible, and what analysis procedures are jointly calibrated with the experimental designs to facilitate the extraction of such information. In recognition of the fact that several excellent book-length treatises exist on the subject matter of design of experiments, this chapter is designed to be only an introduction to some of the most commonly applied techniques, emphasizing principles and, where possible, illustrating practice with relevant examples. Because of the role now played by computer software, our discussion deemphasizes the old-fashioned mechanics of manual computations. This allows us to focus on the essentials of experimental designs and on interpreting the results produced by computer programs.

## 19.1 Introductory Concepts

### 19.1.1 Experimental Studies and Design

In this chapter, the term “experimental studies” is used in a restricted sense to describe those investigations involving the deliberate manipulation of independent variables,  $x_1, x_2, \dots, x_k$ , and observing, by measurement, the response of the dependent variables,  $Y_1, Y_2, \dots, Y_n$ . This is in direct contrast to “observational studies” where the experimenter is a passive observer not directly involved in selecting the values of the independent variables which may (or may not) have been responsible for the observed responses. For example, the earliest studies that led to the hypothesis that smoking may cause lung cancer were based on the observation of higher incidences of lung cancer in smokers compared to non-smokers; the hypothesis was not (and, ethically, could not have been) based on experimental studies where, over a pre-specified period of time, a select group of participants were assigned cigarettes to smoke and the rest were not to smoke. The data were simply “observed,” not “controlled.” On the other hand, in modern clinical studies, the number of groups involved in the study, the assignment of treatment protocols to each group, the collection of data, and every other aspect of the study, are firmly under the control of the experimenters. Observational studies and how to analyze their results are important; but these will not be discussed in this chapter.

To reiterate, the key distinguishing characteristic of the experimental studies of concern in this chapter is that the experimenter is assumed to have control over setting the values of the  $x$  variables at which the experiments are performed. The nature of these studies therefore involves

1. Setting the values of  $x$  where data on  $Y$  is to be acquired, and then
2. Transforming the acquired data (via appropriate analysis and interpretations of the results) to “understanding” as captured via a mathematical representation of the effect of the  $x$  variables on the  $Y$  variables.
3. Such understanding and mathematical representations are then typically used for various applications, ranging from general analysis and prediction to engineered system design, improved system operation, control, and optimization.

The two basic tasks involved in these experimental studies are, therefore:

1. Systematic data acquisition (to ensure informative data sets); and
2. Data transformation to understanding via efficient information extraction.

But carrying out these tasks effectively is complicated by random variation,

because, associated with every observation,  $y_i$ , is an unavoidable and unpredictable fluctuation,  $\epsilon_i$ , masking the true value  $\eta$ ; i.e.,

$$y_i = \eta + \epsilon_i \quad (19.1)$$

As far as the first task is concerned, the key consequence of Eq (19.1) is that for acquired data to be informative, one must find a way to maximize the effect of independent variables on  $y$  and minimize the influence of the random component. With the second task, the repercussion of Eq (19.1) is that efficient data analysis requires using appropriate concepts of probability and statistics presented in earlier chapters. Most revealing, however, is how much Task 1 influences the success of experimental studies: information not contained in the data *cannot* be extracted even by the most sophisticated analysis.

“Statistical Design of Experiments” enables efficient conduct of experimental studies by providing a formal strategy of experimentation and the corresponding analysis technique, jointly calibrated for “optimal” extraction of information in the face of unavoidable random variability. This is especially important when resources, both financial and otherwise, are limited. The techniques to be discussed in this chapter allow the acquisition of richly informative data with the fewest possible experimental runs. Depending on the goals of the investigation — whether it is screening for which independent variables matter the most; developing a quantitative model; finding optimum operating conditions; confirming model predictions, etc — there are appropriate experimental designs specifically calibrated to the task at hand.

### 19.1.2 Phases of Efficient Experimental Studies

Experimental studies that deliver the most benefit for the expended effort are carried out in distinct and identifiable phases:

1. *Planning*: where the scope, and goals of the experimental study are clearly defined;
2. *Design*: where the strategy of experimentation that best fits the goals is determined and the explicit procedure for data gathering is specified;
3. *Implementation*: which involves mostly disciplined execution of the design strategy and the careful collection of the data;
4. *Analysis*: where the appropriate statistics are computed and the results are interpreted.

This chapter is concerned primarily with design and analysis, with the

assumption that the reader/practitioner has been diligent in planning and in implementation.

The explicit layout of specific designs, including the generation of complex designs and the specification of such characteristics as “alias structures” (see later) is now routinely carried out by computer software packages. This is in addition to the computer’s traditional role in relieving the practitioner of the burden of tedious computations involved in data analysis, and in carrying out diagnostic tests. However, while the computer will assist with the design of the experiments and with the analysis of the acquired data, (it might even record the data directly, in some instances), it will not absolve the practitioner of the responsibility to think systematically, and evaluate the result judiciously.

### 19.1.3 Problem Definition and Terminology

The general problem at hand involves testing for the existence of real effects of  $k$  independent variables,  $x_1, x_2, \dots, x_k$ , on possibly several dependent variables. Here are some examples: investigations into the effect of pH ( $x_1$ ), salt concentration ( $x_2$ ), salt type ( $x_3$ ), and temperature ( $x_4$ ), on  $Y$ , the solubility of a particular globular protein; or whether or not different kinds of fertilizers ( $x_1$ : Type I or II), farm location ( $x_2$ : A, B or C), grades of seed ( $x_3$ : Premium, P; Regular, R; or Hybrid, H), show any significant effects on  $Y$ , the yield (in bushels per acre) of soybeans.

The following terminology is standard.

1. *Response*,  $Y$ : is the dependent variable; the objective of the experiment is to measure this response and determine what contributes to its observed value;
2. *Factors*: the independent variables whose effects on the response are to be determined;
3. *Level*: the (possibly qualitative) value of the “factor” employed in the experimental determination of a response; e.g. fertilizer type above has two levels: I and II; farm location has 3 levels, A, B or C; seed grade has 3 levels, P, R, or H.
4. *Treatment*: the various factor-level combinations employed in the experiment, e.g. in the example given above, Fertilizer I, Farm Location A, Seed Grade P constitutes one treatment. This example has  $2 \times 3 \times 3 = 18$  total possible treatments. For a single factor experiment, the levels of this factor are the same as the “treatments,” by default.

## 19.2 Analysis of Variance

Let  $\mu_1, \mu_2, \dots, \mu_k$ , represent the individual “effects” of the  $k$  factors on the response variable,  $Y$ . Then at one level, the objective of most experimental studies, from the simplest to the most complicated, can be summed up as attempts to test the hypothesis that  $\mu_1 = \mu_2 = \dots = \mu_k = 0$  simultaneously; i.e., we presume that these  $k$  factors are all equal in having no “effects” whatsoever on the response  $Y$ , ascribing non-zero effects only if there is sufficient evidence to disprove the null hypothesis.

When  $k > 2$ , this problem is an extension of the testing of equality of two means discussed in Chapter 15. Because of the application of the two-sample  $t$ -test to those problems discussed in Chapter 15, one might be tempted to consider multiple pair-wise  $t$ -tests as a way to handle this current problem of comparing  $k$  means simultaneously. However, such multiple pair-wise comparisons are subject to too high an  $\alpha$ -risk. This is because even when sampling from the same distribution, it can be shown that with multiple pairs, there is a high probability of obtaining — by pure chance alone — differences that are too large.

One effective method for testing hypotheses about the equality of  $k$  population means simultaneously is ANOVA, the analysis of variance technique that made a cameo appearance in a restricted sense in Chapter 16 while we were investigating the significance of the regression model. As we saw briefly during that discussion, ANOVA is predicated on the orthogonal decomposition of the total variation observed in  $Y$  into various constituent components. In the more general treatment, the constituent components are dictated by the problem at hand, which is why with regression, the constituent parts were  $SS_R$ , due to regression, and  $SS_E$ , the left-over residual error sum of squares, as dictated by the regression problem. The contributions of the various components to the total variation in  $Y$  are then tested for significance. Any indication of significance will then suggest non-equality of means.

The two central assumptions in ANOVA are as follows:

**1. Normality and Equal Variance:**

For each treatment, each response,  $Y$ , is a random variable that has a normal distribution with the same (but unknown) variance  $\sigma^2$ . Consequently, factors affect only the mean of  $Y$ , not its variance.

**2. Independence:**

The observations of  $Y$  are independent for each treatment.

The principles behind ANOVA lie at the heart of all the computations

involved in experimental designs, with the nature of the problem at hand naturally determining the design and how the data is analyzed.

The rest of the chapter is now devoted to discussing various experiments and appropriate designs, beginning with the simplest — involving a single factor — and building up to more complex experiments involving multiple factors and complex objectives.

### 19.3 Single Factor Experiments

#### 19.3.1 One-Way Classification

The simplest possible experiment is one involving a single factor and  $k$  levels, giving rise to a total of  $k$  treatments. In this case, the application of treatment  $j$  gives rise to a response  $Y$  with mean  $\mu_j$  and variance  $\sigma^2$ , for all  $j = 1, 2, 3, \dots, k$ .

The data collection is  $k$  sets of random samples of size  $n_1, n_2, \dots, n_j, \dots, n_k$  drawn respectively from populations  $1, 2, \dots, j, \dots, k$ . The primary question of interest is as follows:

Do all treatments have the same effect on the response? i.e., is  
 $\mu_1 = \mu_2 = \dots = \mu_j = \dots = \mu_k ?$

As a concrete example, consider standard size “Raisin Bran” boxes filled on 5 different processing machines, where we are interested in the total number of raisins in a box. More specifically, we are interested in answering the question: “Is there any systematic difference in the number of raisins dispensed per box by the machines?” Does the single factor, “Machine” (of which there are five), have an effect on the response, the number of raisins per box?

Let us now suppose that to answer this question, we choose 6 sample boxes from each machine and count number of raisins in each box. The following is a break down of the characteristics of this problem:

- **Response:** Number of raisins per box;
- **Factor:** 1 (Processing Machine);
- **Levels:** 5 (5 different processing machines);
- **Treatments:** 5;
- **Result:**  $y_{ij}$ , the number of raisins found in box  $i$  filled on machine  $j$ ; this is a realization of the random variable  $Y_{ij}; i = 1, 2, 3, \dots, 6; j = 1, 2, \dots, 5$ ;
- **Assumption:**  $Y_{ij} \sim N(\mu_j, \sigma^2)$ .

### Postulated Model and Hypotheses

For problems of this type, the postulated model is:

$$Y_{ij} = \mu_j + \epsilon_{ij}; i = 1, 2, \dots, n_j; j = 1, 2, \dots, k \quad (19.2)$$

along with the distributional assumption,  $Y_{ij} \sim N(\mu_j, \sigma^2)$ ;  $\mu_j$  is the mean associated with the  $j^{th}$  treatment and  $\epsilon_{ij}$  is the random error. In words, this states that but for the random component, all observations from the  $j^{th}$  treatment are characterized by a constant mean,  $\mu_j$ . The associated hypotheses are:

$$\begin{aligned} H_0 : \quad & \mu_1 = \mu_2 = \dots = \mu_k \\ H_a : \quad & \mu_\ell \neq \mu_m \text{ for some } \ell \text{ and } m. \end{aligned} \quad (19.3)$$

so that all treatments are hypothesized to be identical, unless there is evidence to the contrary.

If we represent as  $\mu$ , the grand mean of the complete data set, and express the  $j^{th}$  treatment mean as:

$$\mu_j = \mu + \tau_j; j = 1, 2, \dots, k \quad (19.4)$$

then  $\tau_j$  represents the  $j^{th}$  treatment effect, that quantity that distinguishes the  $j^{th}$  treatment mean from the grand mean. Observe from Eq (19.4) that by the definition of the grand mean,

$$\sum_{j=1}^k \tau_j = 0 \quad (19.5)$$

(See Exercise 19.2) Furthermore, observe that if the  $j^{th}$  treatment has no effect on  $Y$ ,  $\tau_j = 0$ . As a result, the expression in Eq (19.2) may be rewritten as:

$$Y_{ij} = \mu + \tau_j + \epsilon_{ij}; i = 1, 2, \dots, n_j; j = 1, 2, \dots, k \quad (19.6)$$

and the hypotheses in Eq (19.7) as:

$$\begin{aligned} H_0 : \quad & \tau_1 = \tau_2 = \dots = \tau_k = 0 \\ H_a : \quad & \tau_\ell \neq 0 \text{ for at least one } \ell. \end{aligned} \quad (19.7)$$

### Data Layout and Experimental Design

The layout of the data table from such an experiment is shown in Table 19.1. The appropriate experimental design is the “completely randomized” design where each observation,  $y_{ij}$ , is determined in random order. This is the central feature of this design. Conceptually, it ensures that each  $y_{ij}$  is truly independent of the others, as required by the independence assumption; practically, it has the net impact that any extraneous effects are “broken up”

**TABLE 19.1:** Data table for typical single-factor experiment

| Treatment<br>(Factor level) | 1              | 2              | 3              | ... | $j$            | ... | $k$            |
|-----------------------------|----------------|----------------|----------------|-----|----------------|-----|----------------|
|                             | $y_{11}$       | $y_{12}$       | $y_{13}$       | ... | $y_{1j}$       | ... | $y_{1k}$       |
|                             | $y_{21}$       | $y_{22}$       | $y_{23}$       | ... | $y_{2j}$       | ... | $y_{2k}$       |
|                             | $\vdots$       | $\vdots$       | $\vdots$       | ... | $\vdots$       | ... | $\vdots$       |
|                             | $y_{n_1,1}$    | $y_{n_2,2}$    | $y_{n_3,3}$    | ... | $y_{n_j,j}$    | ... | $y_{n_k,k}$    |
| Total                       | $T_1$          | $T_2$          | $T_3$          | ... | $T_j$          | ... | $T_k$          |
| Means                       | $\bar{y}_{.1}$ | $\bar{y}_{.2}$ | $\bar{y}_{.3}$ | ... | $\bar{y}_{.j}$ | ... | $\bar{y}_{.k}$ |

and distributed evenly among the observations, and not allowed to propagate systematically.

Each treatment is repeated  $n_j$  times, with the repeated observations known as *replicates*. The design is said to be *balanced* if  $n_1 = n_2 = \dots = n_j = \dots = n_k$ ; otherwise, it is *unbalanced*. It can be shown that for a fixed total number of experimental observations,  $N = n_1 + n_2 + \dots + n_j + \dots + n_k$ , the power of the hypothesis test is maximized for the balanced design.

### Analysis

We begin with the following definitions of some averages: the treatment average,

$$\bar{Y}_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij} \quad (19.8)$$

and the grand average,

$$\bar{Y}_{..} = \frac{1}{N} \sum_{j=1}^k \sum_{i=1}^{n_j} Y_{ij}; N = \sum_{j=1}^k n_j \quad (19.9)$$

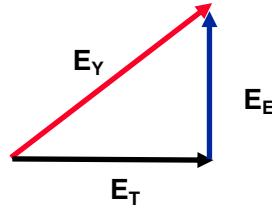
Analyzing the data set in Table 19.1 is predicated on the following data decomposition

$$Y_{ij} = \bar{Y}_{..} + (\bar{Y}_{.j} - \bar{Y}_{..}) + (Y_{ij} - \bar{Y}_{.j}) \quad (19.10)$$

The first term,  $\bar{Y}_{..}$ , is the grand average; the next term,  $(\bar{Y}_{.j} - \bar{Y}_{..})$ , is the deviation from the grand average due to any treatment effect; and the final term is the purely random, within-treatment deviation. (Compare with Eq (19.6).) This expression in Eq (19.10) is easily rearranged to yield:

$$\begin{aligned} (Y_{ij} - \bar{Y}_{..}) &= (\bar{Y}_{.j} - \bar{Y}_{..}) + (Y_{ij} - \bar{Y}_{.j}) \\ E_Y &= E_T + E_E \end{aligned} \quad (19.11)$$

its “error decomposition” version, with the following implications:  $E_Y$ , the vector of deviations of observation  $Y_{ij}$  from the grand average, consists of two



**FIGURE 19.1:** Graphic illustration of the orthogonal vector decomposition of Eq (19.11)

components:  $E_T$ , the component due to any treatment effect, and  $E_E$ , the pure random error component. And now, it turns out that upon taking sums-of-squares in Eq (19.11) the result is the following sums of squares identity (see Exercise 19.3):

$$\begin{aligned} \sum_{j=1}^k \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{..})^2 &= N \sum_{j=1}^k (\bar{Y}_{.j} - \bar{Y}_{..})^2 + \sum_{j=1}^k \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{.j})^2 \\ SS_Y &= SS_T + SS_E \end{aligned} \quad (19.12)$$

which, in terms of vector norms and the definition of error vectors in Eq (19.11), is:

$$\|E_Y\|^2 = \|E_T\|^2 + \|E_E\|^2 \quad (19.13)$$

We now observe that the simultaneous validity of Eqs (19.11) and (19.13) implies that the vectors  $E_T$  and  $E_E$  constitute an orthogonal decomposition of the vector  $E_Y$  as illustrated in Fig 19.1, a multidimensional version of Pythagoras' theorem. This, of course, is the ANOVA identity that we encountered briefly in Chapter 16.

The primary implication of this decomposition for analyzing the single-factor experiment is as follows: whether or not  $H_0$  is true, it can be shown that:

$$E(SS_E) = (N - k)\sigma^2 \quad (19.14)$$

$$E(SS_T) = (k - 1)\sigma^2 + \sum_{j=1}^k n_j \tau_j \quad (19.15)$$

with the following very important consequence: If  $H_0$  is true, then from these two equations, the following mean error sums of squares provide two independent estimates of  $\sigma^2$

$$MS_E = \frac{SS_E}{(N - k)} \quad (19.16)$$

$$MS_T = \frac{SS_T}{(k - 1)} \quad (19.17)$$

**TABLE 19.2:** One-Way Classification ANOVA Table

| Source of Variation       | Degrees of Freedom | Sum of Squares | Mean Square | F           | p |
|---------------------------|--------------------|----------------|-------------|-------------|---|
| Between Treatments        | $k - 1$            | $SS_T$         | $MS_T$      | $MS_T/MS_E$ |   |
| Within Treatments (Error) | $(N - k)$          | $SS_E$         | $MS_E$      |             |   |
| Total                     | $(N - 1)$          | $SS_Y$         |             |             |   |

And now, as a consequence of the normality assumption, the test statistic

$$F = \frac{MS_T}{MS_E} \quad (19.18)$$

possesses an  $F(\nu_1, \nu_2)$  distribution, with  $\nu_1 = k - 1$ , and  $\nu_2 = N - k$ , if  $H_0$  is true. If  $H_0$  is *not* true, as a result of Eq (19.15), the numerator of the  $F$ -statistic will be inflated by  $\sum_{j=1}^k n_j \tau_j / (k - 1)$ .

The analysis of single-factor experiments when carried out using the completely randomized design, is therefore presented in the form of the table shown in Table 19.2, known as a one-way classification ANOVA table. The implied computations are now routinely carried out, and the associated ANOVA tables generated, by computer programs, as illustrated by the next example.

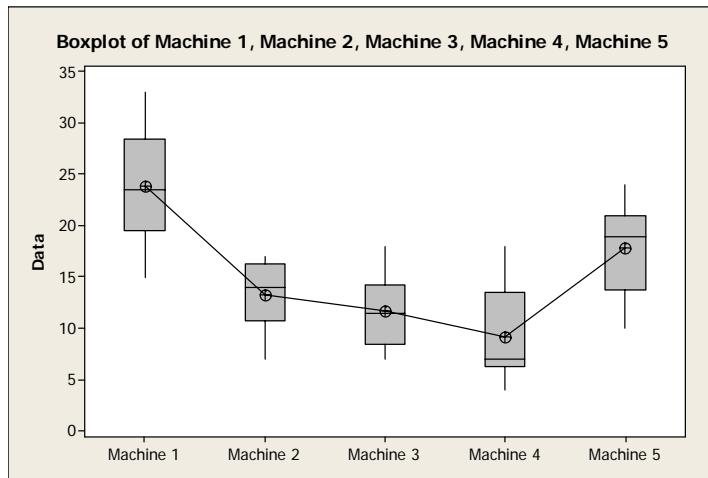
#### **Example 19.1: PERFORMANCE OF RAISIN DISPENSING MACHINES**

Consider the problem of standard size “Raisin Bran” boxes filled on 5 different processing machines, introduced earlier in this chapter. To answer the question posed: “Is there any systematic difference in the number of raisins dispensed per box by the machines?” 6 sample boxes were selected *at random* from each machine, and the number of raisins per box recorded in the table below. Use this data set to answer the question.

|    | Machine 1 | Machine 2 | Machine 3 | Machine 4 | Machine 5 |
|----|-----------|-----------|-----------|-----------|-----------|
| 27 | 17        | 13        | 7         | 15        |           |
| 21 | 12        | 7         | 4         | 19        |           |
| 24 | 14        | 11        | 7         | 19        |           |
| 15 | 7         | 9         | 7         | 24        |           |
| 33 | 14        | 12        | 12        | 10        |           |
| 23 | 16        | 18        | 18        | 20        |           |

#### **Solution:**

To use MINITAB, we begin by entering the provided data into a MINITAB worksheet, RAISINS.MTW; the sequence for carrying out the required analysis is Stat > ANOVA > One-Way (Unstacked). The reason for selecting the “Unstacked” option is that as presented in this



**FIGURE 19.2:** Boxplot of raisins data showing what the ANOVA analysis has confirmed that there is a significant difference in how the machines dispense raisins.

table, the data for each machine is in a different column. This is in contrast to stacking the data for all machines in a single column to which is attached, in another column, the numbers 1, 2, etc, as identifiers for the machine associated with the indicated data.

MINITAB provides several graphical options including box plots for the data, and probability plots for assessing the normality assumption for the residuals. The MINITAB results are summarized below, beginning with the ANOVA table.

**Results for: RAISINS.MTW**  
**One-way ANOVA: Machine 1, Machine 2, Machine 3, Machine 4, Machine 5**

| Source | DF | SS     | MS    | F    | P     |
|--------|----|--------|-------|------|-------|
| Factor | 4  | 803.0  | 200.7 | 9.01 | 0.000 |
| Error  | 25 | 557.2  | 22.3  |      |       |
| Total  | 29 | 1360.2 |       |      |       |

$$S = 4.721 \quad R-Sq = 59.04\% \quad R-Sq(\text{adj}) = 52.48\%$$

The specific value of the  $F$ -statistic, 9.01, indicates that the larger of the two independent estimates of the variance,  $MST$ , the treatment mean squares, is nine times as large as the pure error estimate. It is therefore not surprising that the associated  $p$ -value is 0 to three decimal places. Therefore we must reject the null hypothesis (at the significance level of  $\alpha = 0.05$ ) and conclude that there is a systematic difference in how each machine dispenses raisins. A boxplot of the data is shown in Fig 19.2 for each machine where, visually, we see that Machines 1 and 5 appear to dispense more raisins than the others.

Let us now recall the postulated model for the single-factor experiment in Eq (19.2); treated like a regression model in which 5 parameters  $\mu_j; i = 1, 2, \dots, 5$ , are estimated from the supplied data (representing the mean number of raisins dispensed by Machine  $j$ ), MINITAB provides values for the estimated pure error standard deviation,  $S = 4.72$ , as well as  $R^2$  and  $R_{adj}^2$  values. These have precisely the same meaning as in the regular regression problems discussed in Chapter 16. In addition, the validity of the normality assumptions can be assessed by examining the residuals  $\epsilon_{ij}$ . The normal probability plots for the estimated residuals are shown in Fig 19.3. The top plot is the typical plot obtained directly from the ANOVA dialog in MINITAB; it shows only the residuals and the best normal distribution fit. A visual assessment indicates that the normality assumptions seems to be valid. However, for a more rigorous assessment, MINITAB also provides the option of saving the residuals for further analysis. If this is done, and the rigorous probability model goodness-of-fit test is carried out in conjunction with the probability plot, the result is shown in the bottom panel. The  $p$ -value associated with the A-D test is quite high (0.81) leading us to conclude that the normality assumption indeed appears to be adequate.

### Fixed, Random, and Mixed Effects

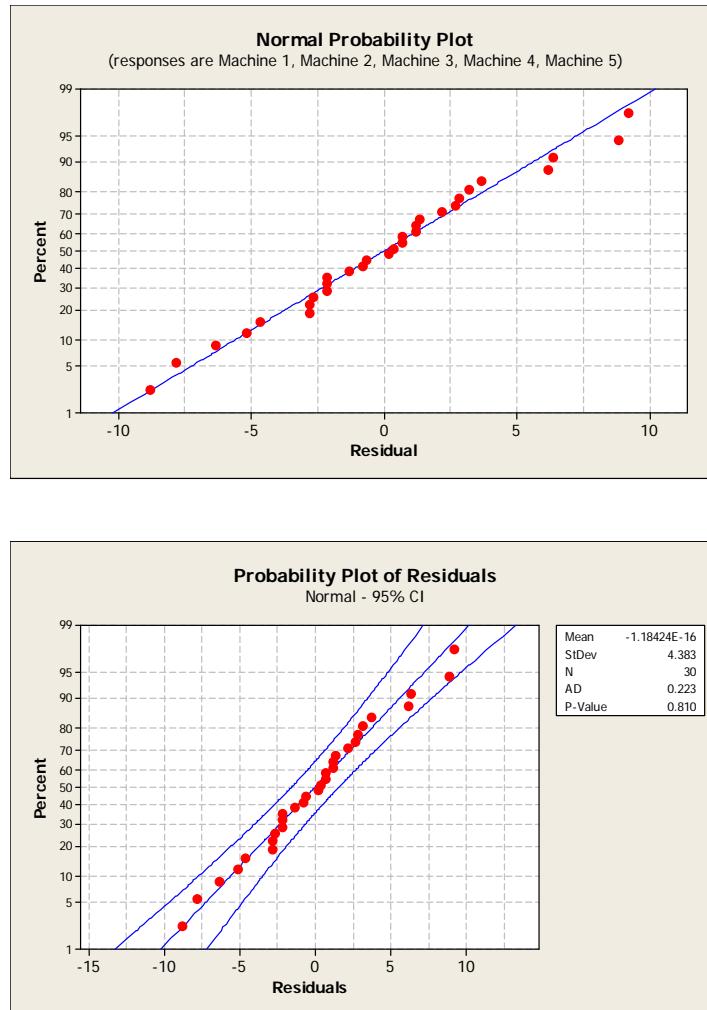
When the  $k$  factor levels (treatments) to be evaluated constitute the complete set of interest, so that the results and conclusions will not be extended to other factor levels that have not been considered explicitly, the problem is considered a “fixed effect” ANOVA problem. The just concluded raisins example posed a fixed effect problem because the 5 machines involved in the experimental study constitute the complete set in the production facility.

On the contrary, when the  $k$  factor levels (treatments) constitute a *random sample* from a larger group of factor level and treatments, and the objective is to extend the results and conclusions to other factor levels that have not been considered explicitly, this is a “random effect” ANOVA problem. If, for example, the 5 machines involved in the raisin example were selected for the experimental study from a complete set of, say, 15 machines in the production facility, that would have turned the problem into a “random effect” ANOVA problem.

With problems involving multiple factors, it is possible for some factors to be fixed while some are random; under these circumstances, the result is a “mixed-effect” ANOVA problem. A discussion of random and mixed-effect ANOVA lies outside the intended scope of this introductory chapter.

### Summary Characteristics

The following are the summary characteristics of the simple, one-way classification single factor experiment:



**FIGURE 19.3:** Normal probability plots of the residuals from the one-way classification ANOVA model in Example 19.1. Top panel: Plot obtained directly from the ANOVA analysis which does not provide any test statistic or significance level; Bottom panel: Subsequent goodness-of-fit test carried out on saved residuals; note the high  $p$ -value associated with the A-D test.

1. *Experiment:* Single Factor ( $k$  levels); fixed or random effect
2. *Design:* Completely randomized; balanced (whenever possible)
3. *Analysis:* One-way ANOVA

### 19.3.2 Kruskal-Wallis Nonparametric Test

In the same spirit as the nonparametric tests discussed in Chapter 18, the Kruskall-Wallis (KW) test is a nonparametric alternative to a one-way ANOVA. Used to test the equality of the *medians* of two or more populations, it is a direct generalization of the strategy underlying the Mann-Whitney-Wilcoxon test, based on the rank of the data values rather than the actual values themselves. As with the other nonparametric tests, the KW test does not make any distributional assumptions (Gaussian or otherwise) about the populations. The only assumption is that the samples are independent random samples from continuous distributions with the same shape.

Thus, when the normality assumption for the one-way ANOVA is invalid, the KW test should be used. Computer programs such as MINITAB provide this option. With MINITAB, the sequence **Stat > Nonparametrics > Kruskall-Wallis >** opens a dialog box similar to the one used for the standard one-way ANOVA.

### 19.3.3 Two-Way Classification

By way of motivation, consider the problem of studying the effect of tire brand on amount of wear experienced by the tire. Specifically, we are interested in answering the question: “Do tire Brands  $A$ ,  $B$ , and  $C$  wear differently?” An experimental investigation of this question might involve outfitting a car with different brands and measuring wear after driving for a pre-specified number of miles. But there are some potential problems: (i) The “Driver Effect”: will the obviously different driving habits of drivers affect the observed wear? This problem may be avoided by using a single driver and then assuming that the driver’s habit will not change over time. (ii) The “Wheel Effect”: the four wheels of a car may not be identical in how they wear down tires.

The challenge is to study “brand effect” while avoiding contamination with the extraneous “wheel effect.” The solution offered by classical design of experiments is to use wheels as a “blocking” variable to obtain the randomized complete block design.

#### The Randomized Complete Block Design

The randomized complete block design, the simplest extension of the completely randomized design, is characterized as follows: it consists of one *primary* factor with  $k$  levels (i.e.,  $k$  treatments) in addition to one blocking vari-

**TABLE 19.3:** Data table for typical single-factor, two-way classification, experiment

| Factor →<br>Blocks ↓ | 1              | 2              | 3              | ... | j              | ... | k              | Means           |
|----------------------|----------------|----------------|----------------|-----|----------------|-----|----------------|-----------------|
| $B_1$                | $y_{11}$       | $y_{12}$       | $y_{13}$       | ... | $y_{1j}$       | ... | $y_{1k}$       | $\bar{y}_{1..}$ |
| $B_2$                | $y_{21}$       | $y_{22}$       | $y_{23}$       | ... | $y_{2j}$       | ... | $y_{2k}$       | $\bar{y}_{2..}$ |
| $\vdots$             | $\vdots$       | $\vdots$       | $\vdots$       | ... | $\vdots$       | ... | $\vdots$       | $\vdots$        |
| $B_r$                | $y_{r1}$       | $y_{r2}$       | $y_{r3}$       | ... | $y_{rj}$       | ... | $y_{rk}$       | $\bar{y}_{r..}$ |
| Total                | $T_1$          | $T_2$          | $T_3$          | ... | $T_j$          | ... | $T_k$          |                 |
| Means                | $\bar{y}_{.1}$ | $\bar{y}_{.2}$ | $\bar{y}_{.3}$ | ... | $\bar{y}_{.j}$ | ... | $\bar{y}_{.k}$ | $\bar{y}_{..}$  |

able variable with  $r$  levels, e.g. wheels 1, 2, 3, 4 in the motivating illustration. The treatments are allocated completely randomly within each block.

In addition to the usual assumptions underlying the completely randomized design, we now also assume that the variability due to the blocking variable is fixed within each block. (For example, what Wheel 1 does to Brand A it does to all other brands.) But this blocking effect, if it exists, may vary from block to block (e.g. the effect of Wheel 1 may be different from that of Wheels 2 or 3 or 4).

### Postulated Model and Hypotheses

For this problem, the postulated model is:

$$Y_{ij} = \mu + \tau_j + \beta_i + \epsilon_{ij}; i = 1, 2, \dots, r; j = 1, 2, \dots, k \quad (19.19)$$

along with the usual distributional assumption,  $Y_{ij} \sim N(\mu_j, \sigma^2)$ ;  $\mu_j$ , the mean associated with the  $j^{th}$  treatment, is the  $j^{th}$  treatment effect;  $\beta_i$  is the  $i^{th}$  block effect and  $\epsilon_{ij}$  is the random error. By definition,

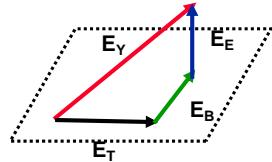
$$\sum_{j=1}^k \tau_j = 0; \sum_{i=1}^r \beta_i = 0 \quad (19.20)$$

Because we are primarily concerned with identifying the presence of a treatment effect, the associated hypotheses are as before in Eq (19.7):

$$\begin{aligned} H_0 : \quad & \tau_1 = \tau_2 = \dots = \tau_k = 0 \\ H_a : \quad & \tau_\ell \neq 0 \text{ for at least one } \ell. \end{aligned}$$

### Data Layout and Experimental Design

The layout of the data table from such an experiment, known as a two-way classification, is shown in Table 19.3. The experimental design is the



**FIGURE 19.4:** Graphic illustration of the orthogonal error decomposition of Eq (19.21) with the additional block component,  $E_B$ .

“randomized complete block” design, two-way crossed, because each factor is combined (crossed) with each block to yield a total of  $r \times k$  experiments; each experimental treatment combination that produces the observation,  $y_{ij}$ , is run in random order.

### Analysis

The analysis technique for this variation on the basic single-factor experiment is similar to the ANOVA technique discussed earlier. This time, however, the orthogonal decomposition has additional terms. It can be shown that in this case, the sum of squares decomposition is

$$\begin{aligned} \|E_Y\|^2 &= \|E_T\|^2 + \|E_B\|^2 + \|E_E\|^2 \\ SS_Y &= SS_T + SS_B + SS_E \end{aligned} \quad (19.21)$$

where, the vector  $E_Y$  of the data deviation from the grand average, is decomposed orthogonally into three components: the vectors  $E_T$ ,  $E_B$  and  $E_E$ , the components due, respectively, to the treatment effect, the block effect, and random error, as illustrated in Fig 19.4. The sum of squares,  $SS_Y$ , the total variability in the data, is thus decomposed into  $SS_T$ , the component due to treatment-to-treatment variability;  $SS_B$  the component due to block-to-block variability; and  $SS_E$ , the component due to pure error. This is a direct extension of the ANOVA identity shown earlier.

It is important to note that with this strategy, the  $SS_B$  component has been separated out from the desired  $SS_T$  component. And now, whether or not  $H_0$  is true, it can be shown that:

$$E(SS_E) = (k - 1)(r - 1)\sigma^2 \quad (19.22)$$

$$E(SS_T) = (k - 1)\sigma^2 + \sum_{j=1}^k \tau_j \quad (19.23)$$

$$E(SS_B) = (r - 1)\sigma^2 + \sum_{i=1}^r \beta_i \quad (19.24)$$

so that if  $H_0$  is true, then from these equations, the following mean error sums

**TABLE 19.4:** Two-Way Classification ANOVA Table

| Source of Variation | Degrees of Freedom      | Sum of Squares | Mean Square | F           | p |
|---------------------|-------------------------|----------------|-------------|-------------|---|
| Between Treatments  | $k - 1$                 | $SS_T$         | $MS_T$      | $MS_T/MS_E$ |   |
| Blocks              | $r - 1$                 | $SS_B$         | $MS_B$      | $MS_B/MS_E$ |   |
| Error               | $(k - 1)*$<br>$(r - 1)$ | $SS_E$         | $MS_E$      |             |   |
| Total               | $(rk - 1)$              | $SS_Y$         |             |             |   |

of squares provide three independent estimates of  $\sigma^2$

$$MS_E = \frac{SS_E}{(k - 1)(r - 1)} \quad (19.25)$$

$$MS_T = \frac{SS_T}{(k - 1)} \quad (19.26)$$

$$MS_B = \frac{SS_B}{(r - 1)} \quad (19.27)$$

Significance is determined using the now-familiar test statistic

$$F = \frac{MS_T}{MS_E} \quad (19.28)$$

which possesses an  $F(\nu_1, \nu_2)$  distribution, this time with  $\nu_1 = k - 1$ , and  $\nu_2 = (k - 1)(r - 1)$ , if  $H_0$  is true. If  $H_0$  is *not* true, from Eq (19.23), the numerator of the F-statistic will be inflated by  $\sum_{j=1}^k \tau_j/(k - 1)$ , but by this term alone. Without separating out the block effect,  $SS_T$  would have been inflated in addition by  $SS_B$ , in which case, even when  $H_0$  is true, the inflationary influence of  $SS_B$  on  $SS_T$  could give the impression that there was a significant treatment effect; i.e., the treatment effect would have been contaminated by the block effect.

The result of the two-way classification, single-factor, randomized block design analysis is presented in the ANOVA table shown in Table 19.4, known as a two-way classification. Once again, this ANOVA table and the computations it entails are routinely generated by computer programs.

**Example 19.2: TIRE WEAR FOR DIFFERENT TIRE BRANDS**

In an experimental study to determine if different brands ( $A, B$  and  $C$ ) of steel-belted radial tires wear differently, a randomized complete block design is used with car wheel as a blocking variable. The problem characteristics are as follows: (i) The response is the amount of wear on the tire (in coded units) after a standardized lab test that amounts to an effective 50,000 road miles; (ii) the factor is “Tire Brand”, (iii) the number of levels is 3 (Brands  $A, B$  and  $C$ ); (iv) the blocking variable

is “Car Wheel” and (v) the number of blocks (or levels) is 4, specifically Wheel 1: Front Left; Wheel 2: Front Right; Wheel 3: Rear Left; and Wheel 4: Rear Right. The data layout is shown in the table below. Determine whether or not the different tire brands wear differently.

| Factor →<br>Blocks ↓ | A    | B     | C     | Means |
|----------------------|------|-------|-------|-------|
| Wheel 1              | 47   | 46    | 52    | 48.33 |
| Wheel 2              | 45   | 43    | 51    | 46.33 |
| Wheel 3              | 42   | 37    | 49    | 42.67 |
| Wheel 4              | 48   | 50    | 55    | 51.00 |
| Means                | 45.5 | 44.00 | 51.75 | 47.08 |

**Solution:**

To use MINITAB, the provided data must be entered into a MINITAB worksheet in stacked format, as shown in the table below:

| Tire Wear | Wheel Number | Tire Brand |
|-----------|--------------|------------|
| 47        | 1            | A          |
| 46        | 1            | B          |
| 52        | 1            | C          |
| 45        | 2            | A          |
| 43        | 2            | B          |
| 51        | 2            | C          |
| 42        | 3            | A          |
| 37        | 3            | B          |
| 49        | 3            | C          |
| 48        | 4            | A          |
| 50        | 4            | B          |
| 55        | 4            | C          |

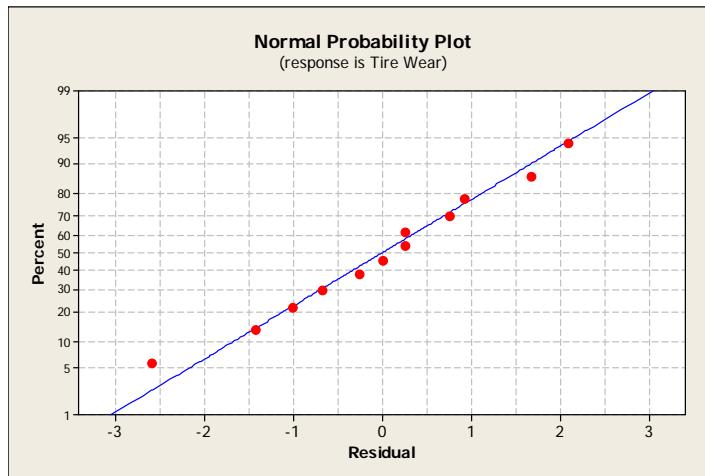
The sequence for carrying out the required analysis is **Stat > ANOVA > Two-Way**, opening a self-explanatory dialog box. The MINITAB results are summarized below.

**Two-way ANOVA: Tire Wear versus Wheel Number, Tire Brand**

| Source       | DF | SS      | MS      | F     | P     |
|--------------|----|---------|---------|-------|-------|
| Wheel Number | 3  | 110.917 | 36.9722 | 11.78 | 0.006 |
| Tire Brand   | 2  | 135.167 | 67.5833 | 21.53 | 0.002 |
| Error        | 6  | 18.833  | 3.1389  |       |       |
| Total        | 11 | 264.917 |         |       |       |

$$S = 1.772 \quad R-Sq = 92.89\% \quad R-Sq(\text{adj}) = 86.97\%$$

The indicated  $p$ -values for both wheel number and tire brand indicate that both effects are significant (at the  $\alpha = 0.05$  level), but the most important fact here is that the wheel effect, has been separated out and distinguished from the primary effect of interest. Therefore we reject the null hypothesis and conclude that there is indeed a systematic difference



**FIGURE 19.5:** Normal probability plots of the residuals from the two-way classification ANOVA model for investigating tire wear, obtained directly from the ANOVA analysis.

in how each tire wears; and we have been able to identify this effect without contamination from the wheel number effect, which is itself also significant.

Again, viewed as a regression model, seven parameters, the 3 treatment effects,  $\tau_j; j = 1, 2, 3$ , and the 4 block effects,  $\beta_i, i = 1, 2, 3, 4$ , can be estimated from the two-way classification model in Eq (19.19). MINITAB provides values for the estimated pure error standard deviation,  $S = 1.772$ ; it also provides values for  $R^2 = 92.89\%$  and  $R_{adj}^2 = 86.97\%$ . These latter values show that a significant amount of the variation in the data has been explained by the two-way classification model. The reduction in the  $R_{adj}^2$  value arises from estimating 7 parameters from a total of 12 data points, with not too many “degrees-of-freedom” left. Still, these values are quite decent.

The normal probability plots for the estimated residuals, obtained directly from the ANOVA dialog in MINITAB, are shown in Fig 19.5. A visual assessment indicates that the normality assumptions appears valid. It is left as an exercise to the reader to carry out the more rigorous assessment by saving the residuals and then carrying out the rigorous probability model goodness-of-fit separately (Exercise 19.7).

It is important to note that in the last example, both “brand effect” and “wheel effect” were found to be significant. Using “wheels” as a blocking variable allowed us to separate out the significant wheel effect from the real object of the investigation; without blocking, the wheel effect would have been compounded with the brand effect. This could have serious repercussions particularly when the primary factor has no effect and the blocking factor has a significant effect.

### 19.3.4 Other Extensions

In the two-way classification case above, the key issue is that in addition to the single factor of interest, there was another variable — a so-called “nuisance” variable — that could potentially contaminate our analysis. In general, with single factor experiments, there is the possibility of more “nuisance” variables, but the approach remains the same: “block” on nuisance variables. When there is only one nuisance variable, the appropriate design, as we have seen, is the randomized complete block design, with analysis provided by the two-way classification ANOVA (one primary factor; one “nuisance” variable). With two blocking variables, the appropriate design is known as the Latin Square design, leading to a three-way classification ANOVA (one primary factor; two “nuisance” variables). With three blocking variables, we use the Graeco-Latin Square design, and a four-way classification ANOVA (one primary factor; three “nuisance variables”) etc. A discussion of these and other such designs lie outside the intended scope of this introductory chapter. (See Box, Hunter and Hunter, 2005<sup>1</sup>)

---

## 19.4 Two-Factor Experiments

With two-factor experiments, there are two legitimate factors of interest (not one factor in conjunction with a “nuisance” variable),  $a$  levels of factor  $A$ , and  $b$  levels of factor  $B$ . Both factors are varied together to produce  $a \times b$  total treatments. For example, consider an experiment in which the effect of temperature and catalyst loading on “conversion” in a batch reactor, is to be investigated at  $150^{\circ}\text{C}$ ,  $200^{\circ}\text{C}$  and  $250^{\circ}\text{C}$  along with 20% and 35% catalyst loading. This is a two-factor experiment, with the “conversion” as the response, and three levels of one factor, temperature, and two levels of the other factor, catalyst loading, for a total of 6 treatments.

The objective in this case is to determine the effects on the response variable of each individual factor, and of the possible interactions between the two factors, when the effect of factor  $A$  changes at different levels of factor  $B$ . It can be shown that ascertaining the interaction effects requires replication of each treatment. The data layout is therefore conceptually similar to that for randomized block case, except in two major ways

1. Now the effect of the second variable is important; and
2. Replication is mandatory.

---

<sup>1</sup>Box, G.E.P., J.S. Hunter, and W.G. Hunter, (2005) *Statistics for Experimenters: Design Innovation and Discovery*, 2nd Ed., Wiley Interscience, N.J.

### Postulated Model and Hypotheses

The postulated model for the two-factor case is:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}; i = 1, 2, \dots, a; j = 1, 2, \dots, b; k = 1, 2, \dots, r \quad (19.29)$$

Here,  $\alpha_i$  is the main effect of the  $i^{th}$  level of Factor A;  $\beta_j$  is the main effect of the  $j^{th}$  level of factor B;  $\gamma_{ij}$  is the effect of the interaction between  $i^{th}$  level of factor A and the  $j^{th}$  level of factor B; and  $\epsilon_{ijk}$  is the random error. Again, by definition,

$$\sum_{i=1}^a \alpha_i = 0; \sum_{j=1}^b \beta_j = 0; \sum_{i=1}^a \sum_{j=1}^b \gamma_{ij} = 0 \quad (19.30)$$

As usual, the distributional assumption is  $Y_{ij} \sim N(\mu_j, \sigma^2)$ . Because we are concerned with identifying all the main effects and interactions, in this case, the null hypotheses are:

$$\begin{aligned} H_0^\alpha : & \alpha_1 = \alpha_2 = \dots = \alpha_a = 0 \\ H_0^\beta : & \beta_1 = \beta_2 = \dots = \beta_b = 0 \\ H_0^\gamma : & \gamma_{ij} = 0; \forall i, j \end{aligned} \quad (19.31)$$

and the alternatives are:

$$\begin{aligned} H_a^\alpha : & \alpha_i \neq 0 \text{ for at least one } i \\ H_a^\beta : & \beta_j \neq 0 \text{ for at least one } j \\ H_a^\gamma : & \gamma_{ij} \neq 0 \text{ for at least one } i, j \text{ pair} \end{aligned} \quad (19.32)$$

### Data Layout and Experimental Design

The layout of the data table for the two-factor experiment is shown in Table 19.5. The experimental design is the “randomized complete block” design, two-way crossed, because each factor is combined (crossed) with each block, and each treatment combination experiment to yield the observation,  $y_{ij}$ , is run in random order.

### Analysis

The analysis for the two-factor problem is based on similar orthogonal data decomposition and sum of squares decomposition; the resulting ANOVA table, consisting of terms for main effects, interaction effects, and error, is shown in Table 19.6.

**TABLE 19.5:** Data table for typical two-factor experiment

| Factor B → | 1         | 2         | ... | b         |
|------------|-----------|-----------|-----|-----------|
| Factor A ↓ |           |           |     |           |
| 1          | $Y_{111}$ | $Y_{121}$ | ... | $Y_{1b1}$ |
| 1          | $Y_{112}$ | $Y_{122}$ | ... | $Y_{1b2}$ |
| :          | :         | :         | ... | :         |
| 1          | $Y_{11r}$ | $Y_{12r}$ | ... | $Y_{1br}$ |
| —          |           |           |     |           |
| 2          | $Y_{211}$ | $Y_{221}$ | ... | $Y_{2b1}$ |
| 2          | $Y_{212}$ | $Y_{222}$ | ... | $Y_{2b2}$ |
| :          | :         | :         | ... | :         |
| 2          | $Y_{21r}$ | $Y_{22r}$ | ... | $Y_{2br}$ |
| —          |           |           |     |           |
| :          | :         | :         | ... | :         |
| a          | $Y_{a11}$ | $Y_{a21}$ | ... | $Y_{ab1}$ |
| a          | $Y_{a12}$ | $Y_{a22}$ | ... | $Y_{ab2}$ |
| :          | :         | :         | ... | :         |
| a          | $Y_{a1r}$ | $Y_{a2r}$ | ... | $Y_{abr}$ |

**TABLE 19.6:** Two-factor ANOVA Table

| Source of Variation     | Degrees of Freedom | Sum of Squares | Mean Square | F              | p |
|-------------------------|--------------------|----------------|-------------|----------------|---|
| Main Effect A           | $a - 1$            | $SS_A$         | $MS_A$      | $MS_A/MS_E$    |   |
| Main Effect B           | $b - 1$            | $SS_B$         | $MS_B$      | $MS_B/MS_E$    |   |
| 2-factor Interaction AB | $(a - 1)*(b - 1)$  | $SS_{AB}$      | $MS_{AB}$   | $MS_{AB}/MS_E$ |   |
| Error                   | $ab(r - 1)$        | $SS_E$         | $SS_E$      |                |   |
| Total                   | $(abr - 1)$        | $SS_Y$         |             |                |   |

## 19.5 General Multi-factor Experiments

The treatment of experiments involving more than two factors is a natural and straightforward extension of just-concluded two-factor discussion. Generally called *factorial* experiments, their distinguishing characteristic is that, as with the two-factor case, the experimental design involves every factor-level combination. Such designs are more precisely referred to as complete factorial designs for the obvious reason that no factor-level combination is left unexplored. But it is this very characteristic that raises a potential problem: as the number of factors increases, the total number of experiments to be performed increases quite dramatically. For example, a case involving 4 factors, with 3 levels each will result in  $3^4 = 81$  total treatments, not counting replicates. With 5 replicates (for determining the various interactions) the number of experiments climbs to 405.

Even if resources were available to perform all these experiments (which is doubtful in many practical circumstances), one of the truly attractive components of modern design of experiments is the underlying philosophy of acquiring the most informative data as judiciously as possible. With multi-factor experiments, therefore, the issue is not so much whether the complete factorial set should be run; the issue is how best to acquire the desired information as judiciously as possible. Sometimes this may mean running the complete factorial set; frequently however, a carefully selected restricted set of treatments can provide sufficiently informative data.

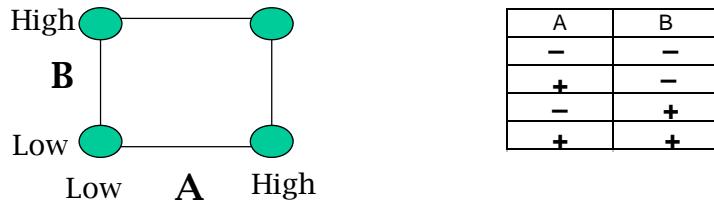
This is the motivation behind one of the most commonly employed designs for multi-factor experiments. When attention is restricted to only two levels of each factor, the result is a special case of general factorial experiments which, for  $k$  factors, gives rise to  $2^k$  observations when all treatments are applied. These are the  $2^k$  factorial designs that are very popular because they are very efficient in how they generate remarkably informative data.

---

## 19.6 $2^k$ Factorial Experiments and Design

### 19.6.1 Overview

$2^k$  factorial designs are used to study the effect of  $k$  factors (and their interactions) simultaneously rather than one-at-a-time. The signature characteristic is that they involve only two levels of each factor, (Low, High). This restriction endows these designs with their key advantage: they are very economical, allowing the extraction of a lot of information with relatively few experiments. There is also the peripheral advantage that the fundamental na-



**FIGURE 19.6:**  $2^2$  factorial design for factors  $A$  and  $B$  showing the four experimental points; – represents low values, + represents high values for each factor.

ture of the design lends itself to computational shortcuts. However, this latter point is no longer of consequence, having been rendered irrelevant by modern computer software. Finally, as we show a bit later,  $2^k$  factorial designs are easily adapted to accommodate experiments involving only a fraction of the total  $2^k$  experiments, especially when  $k$  is so large that even the reduced set of  $2^k$  experiments becomes untenable.

For all their advantages,  $2^k$  factorial experiments also have some important disadvantages. The most obvious is that by restricting attention only to two levels, we limit our ability to confirm the presence of non-linear responses to changes in factor levels. The underlying assumption is that the relationship between the response,  $Y$ , and the factors,  $x_i$ , is approximately linear (plus some possible interaction terms) over the range of the chosen factor levels. When this is a reasonable assumption, nothing is more efficient than  $2^k$  factorial designs. In many practical applications, the recommendation is to use the  $2^k$  factorial designs (or fractions thereof, see later) to begin experimental investigations and then to augment with additional experiments if necessary.

### Notation and Terminology

Because they involve investigations at precisely two levels of each factor, it is customary to use – or –1 to represent the “Low” level and + or +1 to represent the “High” level of each factor. In some publications including journal articles and textbooks, lower case letters  $a, b, c, \dots$  are used to represent the factors, and treatment combinations are represented as: (1),  $a, b, ab, c, ac, bc, abc, \dots$  representing, respectively, the “all low”,  $A$  only high (every other factor low),  $B$  only high,  $A$  and  $B$  high, etc.

For example, a  $2^2$  factorial design involving two factors  $A$  and  $B$  is shown in Fig 19.6. The design calls for a base collection of four experiments, the first, (–, –) representing the (Low, Low) combination; the second, the (High, Low) combination; the third, the (Low, High) combination, and finally, the fourth, the (High, High) combination. A concrete illustration and application of this design is presented in the upcoming Example 19.4.

### Characteristics

The  $2^k$  factorial design enjoys some desirable characteristics that make it particularly computationally attractive:

1. *It is “balanced”*: in the sense that there is an equal number of “highs” and “lows” for each factor. If the factor terms in the design are represented by  $\pm 1$ , then, for each factor,

$$\sum x_i = 0 \quad (19.33)$$

For example, summing down the column for each factor  $A$  and  $B$  in the table on the right hand side of Fig 19.6 shown this clearly.

2. *It is “orthogonal”*: in the sense that the sum of the products of the coded factors (coded as  $\pm 1$ ) is zero, i.e.,

$$\sum x_i x_j = 0; \forall i \neq j \quad (19.34)$$

Multiplying column  $A$  by column  $B$  in Fig 19.6 and summing confirms this for this  $2^2$  example.

These two characteristics simplify analysis, allowing the separation of effects, and making it possible to estimate each effect independently. For example, with the  $2^2$  design of Fig 19.6, from Run #1  $(-, -)$  to Run #2  $(+, -)$ , only factor  $A$  has changed. The difference between the observations,  $y_1$  for Run #1 and  $y_2$  for Run #2, is therefore a reflection of the effect of changing  $A$  (while keeping  $B$  at its low value). The same is true for Runs # 3 and 4, but at the high level of  $B$ ; in which case  $(y_3 - y_4)$  provides another estimate of the main effect of factor  $A$ . This main effect is therefore estimated from the results of the  $2^2$  design as:

$$\text{Main Effect of } A = \frac{1}{2} [(y_2 - y_1) + (y_3 - y_4)] \quad (19.35)$$

The other main effect and the two-way interaction can be computed from similar considerations made possible by the transparency of the design.

When experimental data were analyzed by hand, these characteristics led to the development of useful computational shortcuts (for example the popular Yates’ algorithm); this is no longer necessary, because of computer software packages.

#### 19.6.2 Design and Analysis

The general postulated model for  $2^k$  factorial designs is:

$$y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \sum_{i=1}^k \sum_{j=1}^k \beta_{ij} x_i x_j + \sum_{i=1}^k \sum_{j=1}^k \sum_{\ell=1}^k \beta_{ij\ell} x_i x_j x_\ell + \cdots + \epsilon \quad (19.36)$$

a somewhat unwieldy looking equation that is actually quite simple for specific cases. The parameter  $\beta_0$  represents the grand average;  $\beta_i$  is the coefficient related to the main effect of factor  $i$ ; the double subscripted parameter  $\beta_{ij}$  is related to the two-way interaction effect of the  $i^{th}$  and  $j^{th}$  factors; the triple subscripted parameter  $\beta_{ij\ell}$  is related to the three-way interaction effect of the  $i^{th}$ ,  $j^{th}$  and  $\ell^{th}$  factors, etc. Simpler, specific cases are illustrated with the next example.

**Example 19.3: TWO-FACTOR AND THREE-FACTOR  $2^k$  FACTORIAL MODELS**

Write the postulated  $2^k$  factorial models for the following two practical experimental cases:

- (1) Studying the growth of epitaxial layer on silicon wafer by chemical vapor deposition (CVD), where the response of interest, the epitaxial layer thickness,  $Y$ , is believed to depend on two primary factors: deposition time,  $x_1$ , and arsenic flow rate,  $x_2$ .
- (2) Characterizing the solubility,  $Y$ , of a globular protein as a function of three primary factors: pH,  $x_1$ ; salt concentration,  $x_2$ ; and temperature,  $x_3$ .

**Solution:**

The models are as follows: for the CVD experiment,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon \quad (19.37)$$

$\beta_1$  is related to the main effect of deposition time on epitaxial layer thickness;  $\beta_2$  to the main effect of arsenic flow rate; and  $\beta_{12}$  is the single two-way interaction coefficient representing how deposition time interacts with arsenic flow rate.

For the globular protein solubility experiment, the model is:

$$\begin{aligned} y = & \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 \\ & + \beta_{23} x_2 x_3 + \beta_{123} x_1 x_2 x_3 + \epsilon \end{aligned} \quad (19.38)$$

The objective is to collect data according to the  $2^k$  factorial design and determine values for the indicated parameters by analyzing the data. Because of the design, all the main effects and interactions are in fact estimated from arithmetic means. Of course, not all parameters will be “important;” and, as usual, which one is important and which one is not is determined from tests of significance that are based on the same ANOVA decomposition presented above for the two-factor experiment, as well as on individual  $t$ -tests for the parameter estimates. The null hypothesis is that all these parameters are identically equal to 0; significance is determined usually at the  $\alpha = 0.05$  level.

### 19.6.3 Procedure

The general procedure for carrying out  $2^k$  factorial experiments is summarized here:

1. *Create Design:*

Given the number of factors, and the low and high levels for each factor, use computer software (e.g. MINITAB, SAS, ...) to create the design. Using these software packages to create the design is straightforward and intuitive; the result is a design table showing each treatment combination and the recommended run sequence. It is highly recommended to run the experiments in random order, just as with the completely randomized single-factor designs.

Creating the design requires determining how many replicates to run. We present some recommendations shortly.

2. *Perform Experiment:*

This involves filling in the results of each experiment in the data sheet created above.

3. *Analyze Data:*

Once the data is entered into the created design worksheet, all statistical packages will carry out the factorial analyses, generating values for the “main effects” and interactions, and providing associated significance levels. It is also recommended to carry out diagnostic checks to validate the underlying Gaussian distributional assumption.

### Sample Size Considerations

The discussion in Section 15.5 of Chapter 15 included practical considerations for determining sample sizes for carrying out hypothesis tests regarding means of normal populations. The same considerations can be extended for balanced 2-level factorial designs. In the same spirit, let  $\delta^*$  represent the smallest “shift” from zero we wish to detect in these effects (i.e., the smallest magnitude of the effect worth detecting). With  $\sigma$ , the standard deviation of the random error component,  $\epsilon$ , usually unknown ahead of time, we can invoke the definition of the “signal-to-noise” ratio given in Chapter 15, i.e.,

$$\rho_{SN} = \frac{\delta^*}{\sigma} \quad (19.39)$$

and use it in conjunction with the following expression to determine a range of recommended sample sizes (rounded up to the nearest integer).

$$\left( \frac{7}{\rho_{SN}} \right)^2 < n < \left( \frac{8}{\rho_{SN}} \right)^2 \quad (19.40)$$

(Compare with Eq (15.95).) The following table generated from Eq (19.40) is roughly equivalent to Table 15.8.

| $\rho_{SN}$ | 0.5     | 1.0   | 1.5   | 2.0   |
|-------------|---------|-------|-------|-------|
| $n$         | 196-256 | 49-64 | 22-28 | 12-16 |

Alternatively, the sequence, **Stat > Power and Sample Size > 2-level Factorial Design** in MINITAB will also provide recommended values.

The following example illustrates these principles.

**Example 19.4:  $2^2$  FACTORIAL INVESTIGATION OF CVD PROCESS<sup>2</sup>**

In an experimental study of the growth of epitaxial layer on silicon wafer by chemical vapor deposition (CVD), the effect of deposition time and arsenic flow rate on the epitaxial layer thickness was studied in a  $2^2$  factorial experiment using the following settings:

- Factor A: Deposition time (Short, Long)
- Factor B: Arsenic Flow Rate (55%; 59%)

Designing for a signal-to-noise ratio of 2 leads to a recommended sample size  $12 < n < 16$ , which calls for a minimum of three full replicates of the four basic  $2^2$  experiments (and a maximum of four replicates). The  $2^2$  factorial design, in 3 replicates, and the experimental results are shown in the table below. Analyze the data and comment on the results.

| RunOrder | Depo.<br>Time | Arsenic<br>Flow rate | Epitaxial<br>Layer Thick. |
|----------|---------------|----------------------|---------------------------|
| 1        | -1            | -1                   | 14.04                     |
| 2        | 1             | -1                   | 14.82                     |
| 3        | -1            | 1                    | 13.88                     |
| 4        | 1             | 1                    | 14.88                     |
| 5        | -1            | -1                   | 14.17                     |
| 6        | 1             | -1                   | 14.76                     |
| 7        | -1            | 1                    | 13.86                     |
| 8        | 1             | 1                    | 14.92                     |
| 9        | -1            | -1                   | 13.97                     |
| 10       | 1             | -1                   | 14.84                     |
| 11       | -1            | 1                    | 14.03                     |
| 12       | 1             | 1                    | 14.42                     |

**Solution:**

First, to create the design using MINITAB, the required sequence is **Stat > DOE > Factorial > Create Factorial Design**, which opens a dialog box where all the characteristics of the problem are specified. The result, which should be stored in a worksheet, is the set of 12 treatment combinations arising from 3 replicates of the base  $2^2$  design. It is highly recommended to run the actual experiments in random order. (MINITAB can provide a randomized order for the experiments upon selecting the “Randomize runs” option.)

Once the experiments have been completed and the epitaxial layer thickness measurements entered into the worksheet (as shown in the table above), the sequence for analyzing the data is **Stat > DOE > Factorial > Analyze Factorial Design**. The dialog box contains

<sup>2</sup>Adapted from an article in the AT & T Technical Journal

many options and the reader is encouraged to explore them all. The basic MINITAB output is shown here.

**Factorial Fit: Epitaxial Layer versus Depo. Time, Arsenic Flow rate**

**Estimated Effects and Coefficients for Epitaxial Layer Thick. (coded units)**

| Term              | Effect | Coef    | SE Coef | T       | P     |       |
|-------------------|--------|---------|---------|---------|-------|-------|
| Constant          |        | 14.3825 | 0.04515 | 318.52  | 0.000 |       |
| Depo. Time        |        | 0.7817  | 0.3908  | 0.04515 | 8.66  | 0.000 |
| Arsenic Flow rate |        | -0.1017 | -0.0508 | 0.04515 | -1.13 | 0.293 |
| Depo. Time*       |        | 0.0350  | 0.0175  | 0.04515 | 0.39  | 0.708 |
| Arsenic Flow rate |        |         |         |         |       |       |

$$S = 0.156418 \quad R-Sq = 90.51\% \quad R-Sq(adj) = 86.96\%$$

This MINITAB output lists both the “coefficient” and “effect” associated with each factor and it should be obvious that one is twice the other. The “coefficient” represents the parameters in the factorial model equation, e.g. as in Eq (19.37) and (19.38); from these equations, we see that each coefficient represents the change in  $y$  for a *unit change* in each factor  $x_i$ . On the other hand, what is known as the “effect” is the change in  $y$  when each factor,  $x_i$ , changes from its low value to its high value; i.e., from  $-1$  to  $+1$ . Since this is a change of two units in  $x$ , the effects are therefore twice the magnitude of the coefficients. The constant term is unaffected since it is not multiplied by any factor; it is the grand average of the data.

Next, we observe that for each estimated parameter, there is an associated *t*-statistic and corresponding *p*-value. These are determined in the same manner as discussed in earlier chapters, using the Gaussian assumption and the mean square error,  $MS_E$ , as an estimator of  $S$ , the standard deviation. The null hypothesis is that all coefficients are identically zero. In this specific case, only the constant term and the Deposition time effect appear to be significantly different from zero; the Arsenic flow rate term and the two-factor interaction term appear not to be significant at the  $\alpha = 0.05$  level. MINITAB also lists the usual regression characteristics,  $R^2$ ;  $R_{adj}^2$  none of which appear to be particularly indicative of anything unusual.

Finally, MINITAB also produces an ANOVA table shown here. This is a consolidation of the detailed information already shown above. It indicates that the main effects, as a composite (not separating one main effect from the other) are significant (*p*-value is 0.000); the two-way interaction is not (*p*-value is 0.708).

**Analysis of Variance for Epitaxial Layer Thick. (coded units)**

| Source             | DF | Seq SS  | Adj SS  | Adj MS   | F     | P     |
|--------------------|----|---------|---------|----------|-------|-------|
| Main Effects       | 2  | 1.86402 | 1.86402 | 0.932008 | 38.09 | 0.000 |
| 2-Way Interactions | 1  | 0.00368 | 0.00368 | 0.003675 | 0.15  | 0.708 |
| Residual Error     | 8  | 0.19573 | 0.19573 | 0.024467 |       |       |
| Total              | 11 | 2.06343 |         |          |       |       |

Thus, the final result of the analysis is that if  $x_1$  represents deposition time, and  $x_2$ , arsenic flow rate, then the estimated equation relating these factors to  $y$ , the epitaxial layer thickness, is

$$y = 14.3825 + 0.3908x_1 \quad (19.41)$$

where, it must be kept in mind, that  $x_1$  is coded as  $-1$  for the “short” deposition time, and  $+1$  for the “long” deposition time.

#### 19.6.4 Closing Remarks

The best applications of  $2^k$  factorial designs are for problems with relatively small number of factors, say  $k < 5$ ; and when the relationship between  $y$  and the factors is reasonably linear in the “Low-High” range explored experimentally, with non-linearities limited to cross-terms in the factors (no quadratic or higher order effects). The direct corollary is that practical problems not suited to  $2^k$  factorial designs include those with a large number of potentially important factors; and those for which nonlinear relationships are important to the investigation, for example, for applications in process optimization.

For the latter group of problems, specific extensions of the  $2^k$  factorial designs are available to deal with each problem:

1. Screening Designs (for a large number of factors);
2. Response Surface Methods (for more complex modeling and optimization).

We deal next with these designs in turn, beginning with screening designs.

Screening designs are experimental designs specifically calibrated for selecting from among a large group of potential factors, only the few that are truly important, prior to carrying out more detailed experiments. Such designs therefore involve significantly fewer experimental runs than required by the  $2^k$  factorial designs. They are created to avoid the expenditure of a lot of experimental effort since the objective is quicker decisions on factor importance, rather than detailed characterization of effects on responses. The two most popular categories are Fractional Factorial Designs and Plackett-Burman designs.

While screening designs involve running fewer experiments than called for by the  $2^k$  factorial designs, response surface designs involve judiciously adding more experimental runs, to capture more complex relationships better.

## 19.7 Screening Designs: Fractional Factorial

### 19.7.1 Rationale

Investigations with large  $k$  require an exponential increase in the total number of experimental runs if the full  $2^k$  factorial design is employed. With an increase in the number of factors also comes an increase in the total number of higher-order interactions. For example, for the case with 5 factors, with a  $2^5$  factorial design and the resulting base 32 experimental results, we are able to estimate 1 overall average; 5 main effects; 10 2-factor interactions ( $5 \times 4/2$ ); 10 3-factor interactions ( $5 \times 4/2$ ); 5 4-factor interactions; and 1 5-factor interaction. This raises an important practical question: *How important or even physically meaningful are fourth- and higher-order interactions?* The answer is that in all likelihood, they will be either unimportant or at best insignificant. The underlying rationale behind fractional factorial designs is to give up ability to estimate (unimportant) higher order interaction effects in return for fewer experimental runs when  $k > 5$

### 19.7.2 Illustrating the Mechanics

Consider a case involving 4 factors,  $A, B, C, D$ , for which a full  $2^4$  factorial design will require 16 base experiments (not counting replicates). Let us investigate the possibility of making do with *half* of the experiments, based on a  $2^3$  factorial design for the first 3 factors  $A, B, C$ . This proposition may be possible if we are willing to give up something in this reduced design to be used for the other factor  $D$ . First, the base  $2^3$  factorial design for  $A, B$  and  $C$  is shown here:

| Run # | A  | B  | C  |
|-------|----|----|----|
| 1     | -1 | -1 | -1 |
| 2     | 1  | -1 | -1 |
| 3     | -1 | 1  | -1 |
| 4     | 1  | 1  | -1 |
| 5     | -1 | -1 | 1  |
| 6     | 1  | -1 | 1  |
| 7     | -1 | 1  | 1  |
| 8     | 1  | 1  | 1  |

Now suppose that we are willing to give up the ability to determine the three-way interaction  $ABC$ , in return for being able to investigate the effect of  $D$  also. In the language of fractional factorial design, this is represented as:

$$D = ABC \quad (19.42)$$

an expression to which we shall return shortly. It can be shown that the code corresponding to  $ABC$  is obtained from a term-by-term multiplication of the

signs in the columns  $A$ ,  $B$ , and  $C$  in the design table. Thus, for example, the first entry for run #1 will be  $-1$  ( $= -1 \times -1 \times -1$ ), run #2 will be  $1$  ( $= 1 \times -1 \times -1$ ), etc. The result is the updated table shown here.

| Run # | A  | B  | C  | D  |
|-------|----|----|----|----|
| 1     | -1 | -1 | -1 | -1 |
| 2     | 1  | -1 | -1 | 1  |
| 3     | -1 | 1  | -1 | 1  |
| 4     | 1  | 1  | -1 | -1 |
| 5     | -1 | -1 | 1  | 1  |
| 6     | 1  | -1 | 1  | -1 |
| 7     | -1 | 1  | 1  | -1 |
| 8     | 1  | 1  | 1  | 1  |

Thus, by giving up the ability to estimate the three-way interaction  $ABC$ , we have obtained the 8-run design shown in this table for 4 factors, a 50% reduction in the number of runs (i.e., a half-fraction of what should have been a full  $2^4$  factorial design). This seems like a reasonable price to pay; however, this is not the whole cost. Observe that the code for the two-way interaction  $AD$  (obtained by multiplying each term in the  $A$  and  $D$  columns) written horizontally, is  $(1, 1, -1, -1, -1, -1, 1, 1)$  which is precisely the same as the code for the two-way interaction  $BC$ ! But even that is still not all. It is left as an exercise to the reader to show that the code for  $AB$  is also the same as that for  $CD$ ; similarly, the codes for  $AC$  and for  $BD$  are also identical.

Observe therefore that for this problem,

1. The *primary* trade-off,  $D = ABC$ , allowed an eight-run experimental design for a 4-factor system, precisely half of the 16 runs ordinarily required for a full  $2^4$  design for 4 factors;
2. But  $D = ABC$  is not the only resulting trade-off; other trade-offs include  $AD = BC$ ,  $AB = CD$ ; and  $AC = BD$ ; plus some others;
3. The implication of these secondary (collateral) trade-offs is that these two-way interactions, for example,  $AD$  and  $BC$ , are now “confounded”, being indistinguishable from each other; they cannot be estimated independently.

Thus, when we give up some high-order interactions to estimate some other factors, we also lose the ability to estimate other additional effects independently.

### 19.7.3 General characteristics

#### Notation and Alias Structure

The illustrative example 8-run factorial design for a 4-factor system shown above is a half-factorial of a  $2^4$  design, called a  $2^{4-1}$  design. In general a  $2^{k-p}$

design is a  $(2^{-p})$  fraction of the full  $2^k$  design. For example, a  $2^{5-2}$  design is a quarter fraction of the full  $2^5$  design which consists of 8 total runs (1/4 of the full 32).

As illustrated above, the reduction in the total number of runs in  $2^{k-p}$  designs is achieved at a cost; this cost of “fractionation,” the confounding of two effects so that they cannot be independently estimated, is known as “Aliasing.” And for every fractional factorial design, there is an accompanying “alias structure,” a complete listing of what is confounded with what. Such alias structures can be determined from what is known as the *defining relation*, an expression, such as the one in Eq (19.42) above, indicating the primary trade-off.

There are simple algebraic rules for determining alias structures. For instance, upon “multiplying” both sides of Eq (19.42) by  $D$  and using the simple rule that  $DD = I$ , the identity column, we obtain,

$$I = ABCD \quad (19.43)$$

This is the defining relation for this particular fractional factorial design. The additional aliases can be obtained using the same algebraic rule: upon multiplying both sides of Eq (19.43) by  $A$ , and then by  $B$ , and then  $C$ , we obtain:

$$A = BCD; B = ACD; C = ABD \quad (19.44)$$

showing that, like the main effect  $D$ , the other main effects  $A$ ,  $B$ , and  $C$  are also confounded with the indicated three-way interactions. From here, upon multiplying the expressions in Eq (19.44) by the appropriate letters, we obtain the following additional aliases:

$$AB = CD; AC = BD; AD = BC \quad (19.45)$$

Observe that for this design, main effects are confounded with 3-way interactions only; and 2-way interactions are confounded with other 2-way interactions.

### Design Resolution

The order of the effects that are aliased is captured succinctly in the “design resolution.” For example, the illustration used above, is a “Resolution IV” design because 2-way interactions are confounded with other 2-way interactions, and main effects (“1-way interactions”) are confounded with 3-way interactions.

The resolution of a design is typically represented by roman numerals, III, IV, V, etc.; they define the cost of “fractionation.” The higher the design resolution, the better we are able to determine main effects, two-way interactions (and even 3-way interactions) independently. The following notation is typical:  $2_{IV}^{4-1}$  represents a Resolution IV, half fraction of a  $2^4$  factorial design, such as the illustrative example used above. On the other hand, a  $2_{III}^{5-2}$  design is a Resolution III quarter fraction of a  $2^5$  design.

In general, an  $m$ -way interaction aliased with an  $n$ -way interaction implies a Resolution  $(m + n)$ . The following are some general remarks about design resolutions:

1. **Resolution III:** Main effects (1-way) are aliased with 2-way interactions. Use only sparingly.
2. **Resolution IV:** Main effects (1-way) aliased with 3-way interactions; and 2-way interactions aliased with 2-way interactions. This is usually acceptable for many applications.
3. **Resolution V:** Main effects (1-way) aliased with 4-way interactions; 2-way interactions aliased with 3-way interactions. This is usually an excellent choice whenever possible. Because one is usually not so much concerned about 3- or higher-order interactions, main effects and 2-way interactions that are of importance can be estimated almost “cost-free”.

Designs with resolution higher than V are economical, virtually “cost-free” means of estimating main effects and 2-way interactions.

#### 19.7.4 Design and Analysis

##### Basic Principles

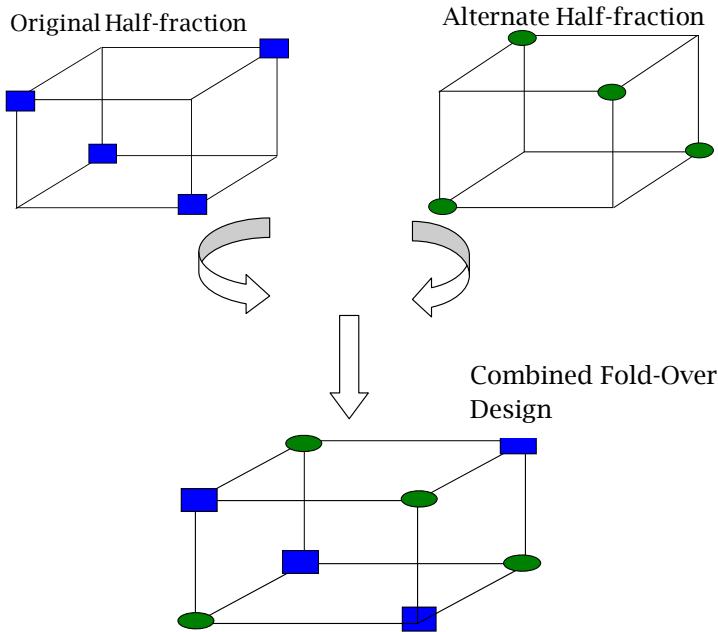
The design and analysis of fractional factorial designs are very similar to those for the basic factorial designs. Here are the key points of interest.

1. *Planning:* Determine  $k$  (total number of factors), and  $p$  (extent of fractionation), giving the total number of unreplicated runs; determine how many replicates are required using the rule-of-thumb specified above or the MINITAB “power and sample size” feature.
2. *Design:* Given the above information, typical computer software packages will display available designs and their respective resolutions; select appropriate resolution (recommend IV or higher where possible). The computer program will then generate the design and the accompanying alias structure. It is important to save both into a worksheet.
3. *Analysis:* This follows the same principles as the full  $2^k$  designs, but in interpreting the results, keep aliases in mind.

Before dealing with a comprehensive example, there are two more concepts of importance in engineering practice we wish to discuss.

##### Projection and Folding

Consider, for example, that one or more factors in a half fraction of a  $2^k$  design (i.e., a  $2^{k-1}$  design) is not significant, we can then “project” the data



**FIGURE 19.7:** Graphic illustration of “folding” where two half-fractions of a  $2^3$  factorial design are combined to recover the full factorial design; each fold costs an additional degree of freedom for analysis.

down to a full factorial in the significant  $k - 1$  variables. For instance, imagine that after completing the experiments in the 4-factor illustrative example presented above, we discover that  $D$  was not significant after all; observe that the data can then be considered as arising from a full  $2^3$  factorial design in  $A, B$  and  $C$ . If, for example, both  $C$  and  $D$  are not significant, then the data can be projected down two full dimensions so that the 8 experiments will be considered as 2 full replicates of a full  $2^2$  factorial design in the significant factors  $A$  and  $B$ . Everywhere the factor  $C$  was investigated therefore becomes a replicate. Where the opportunity presents itself, projection therefore always strengthens the experimental analysis in the remaining factors. This is illustrated shortly with a concrete example.

The reverse is the case with “folding” — combining lower fraction designs into higher fraction ones. For example, combining two  $1/2$  fractions into a full factorial design, or two  $1/4$  fractions into a  $1/2$  fractional factorial design. This is illustrated in Fig 19.7. This strategy is employed if, after the analysis of the fractional factorial design results, we discover that some of the confounded interactions and effects are important enough to be determined independently. Folding increases the design resolution by providing additional information required to resolve aliases. However, each fold costs an additional degree of freedom for analysis.

While we are unable to accommodate additional detailed discussions of fractional factorial designs in such an introductory chapter as this, we use the following example to illustrate its practical application.

### 19.7.5 A Practical Illustrative Example

#### Problem Statement

The problem involves a single-wafer plasma etcher process which uses the reactant gas,  $C_2F_6$ . The response of interest is the etch rate for Silicon Nitride (in Å/min), which is believed to be dependent on the 4 factors listed below.

- A: Gap; cm. (The spacing between the Anode and Cathode)
- B: Reactor Chamber Pressure; mTorr.
- C: Reactant Gas ( $C_2F_6$ ) Flow rate; SCCM
- D: Power (applied to Cathode); Watts.

The objective is to determine which factors affect etch rate by investigating the process response at the values indicated below for the factors using a  $2^{4-1}$  design with no replicates.

| Variable                | Levels |     |
|-------------------------|--------|-----|
|                         | -1     | +1  |
| A. Gap (cm)             | 0.8    | 1.2 |
| B. Pressure (mTorr)     | 450    | 550 |
| C. Gas Flow Rate (sccm) | 125    | 200 |
| D. Power (Watts)        | 275    | 325 |

#### Design and Data Collection

The required design is created in MINITAB using the sequence **Stat > DOE > Factorial > Create Factorial Design** and entering the problem characteristics. MINITAB returns both the design (which is saved into a worksheet) and the following characteristics, including the alias structure.

#### Fractional Factorial Design

|          |      |                     |      |             |     |
|----------|------|---------------------|------|-------------|-----|
| Factors: | 4    | Base Design:        | 4, 8 | Resolution: | IV  |
| Runs:    | 8    | Replicates:         | 1    | Fraction:   | 1/2 |
| Blocks:  | none | Center pts (total): | 0    |             |     |

Design Generators: D = ABC

Alias Structure

I + ABCD

A + BCD

B + ACD  
 C + ABD  
 D + ABC  
 AB + CD  
 AC + BD  
 AD + BC

Note that this is precisely the same design generator and alias structure as in the Resolution IV example used above to illustrate the mechanics of fractional factorial design generation.

The design table, along with the data acquired using the design are shown below:

| Std Order | Run Order | Gap | Pressure | Gas Flow | Power | Etch Rate |
|-----------|-----------|-----|----------|----------|-------|-----------|
| 1         | 5         | 0.8 | 450      | 125      | 275   | 550       |
| 2         | 7         | 1.2 | 450      | 125      | 325   | 749       |
| 3         | 1         | 0.8 | 550      | 125      | 325   | 1052      |
| 4         | 8         | 1.2 | 550      | 125      | 275   | 650       |
| 5         | 6         | 0.8 | 450      | 200      | 325   | 1075      |
| 6         | 4         | 1.2 | 450      | 200      | 275   | 642       |
| 7         | 2         | 0.8 | 550      | 200      | 275   | 601       |
| 8         | 3         | 1.2 | 550      | 200      | 325   | 729       |

Note the difference between the randomized order in which the experiments were performed and the standard order.

### Analysis Part 1

To analyze this data set, the sequence **Stat > DOE > Factorial > Analyze Factorial Design >** opens a dialog box with several self-explanatory options. Of these, we draw particular attention to the button labeled "Terms." Upon selecting this button, a further dialog box is opened in which the terms to be included in the analysis are shown. It is interesting to note that the default already selected by MINITAB shows only the four main effects, *A, B, C, D* and three two way interactions *AB, AC* and *AD*. The reason, of course, is that everything else is aliased with these terms. Next, the button labeled "Plots" allows one to select which plots to display. For reasons that will become clearer later, we select for the "Effect Plots" the "Normal Plot" option. The button labeled "Results" shows what MINITAB will include in the output — estimated coefficients and ANOVA table, Alias table with default interactions, etc. The results for this particular analysis are shown here.

### Results for: Plasma.MTW

Factorial Fit: Etch Rate versus Gap, Pressure, Gas Flow, Power

Estimated Effects and Coefficients for Etch Rate (coded units)

| Term         | Effect  | Coef   |
|--------------|---------|--------|
| Constant     | 756.00  |        |
| Gap          | -127.00 | -63.50 |
| Pressure     | 4.00    | 2.00   |
| Gas Flow     | 11.50   | 5.75   |
| Power        | 290.50  | 145.25 |
| Gap*Pressure | -10.00  | -5.00  |
| Gap*Gas Flow | -25.50  | -12.75 |
| Gap*Power    | -197.50 | -98.75 |

S = \* PRESS = \*

## Alias Structure

I + Gap\*Pressure\*Gas Flow\*Power  
 Gap + Pressure\*Gas Flow\*Power  
 Pressure + Gap\*Gas Flow\*Power  
 Gas Flow + Gap\*Pressure\*Power  
 Power + Gap\*Pressure\*Gas Flow  
 Gap\*Pressure + Gas Flow\*Power  
 Gap\*Gas Flow + Pressure\*Power  
 Gap\*Power + Pressure\*Gas Flow

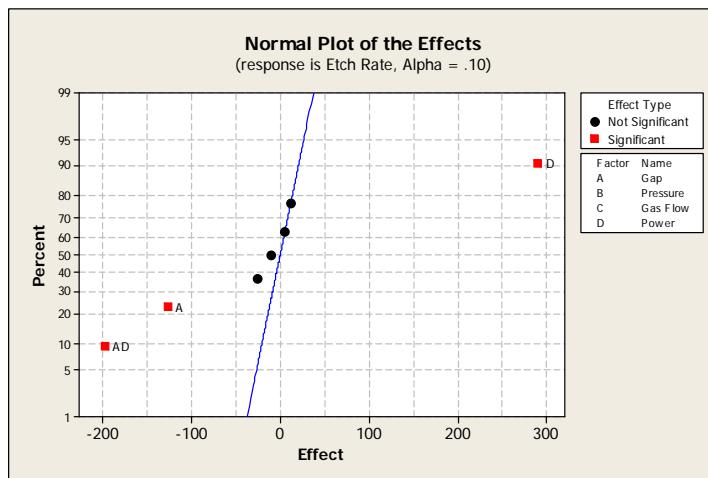
Analysis of Variance for Etch Rate (coded units)

| Source             | DF | Seq SS | Adj SS | Adj MS | F | P |
|--------------------|----|--------|--------|--------|---|---|
| Main Effects       | 4  | 201335 | 201335 | 50334  | * | * |
| 2-Way Interactions | 3  | 79513  | 79513  | 26504  | * | * |
| Residual Error     | 0  | *      | *      | *      |   |   |
| Total              | 7  | 280848 |        |        |   |   |

The first thing we notice is that the usual  $p$ -values associated with the estimates are missing in both the “Estimated Effects” table, and in the ANOVA table. Because there are no replicates, there is no independent way to estimate the random error component of the data, from which the standard deviation,  $s$ , is estimated, which is in turn used to determine significance ( $p$ -values). This is why no value is returned for  $S$  by MINITAB.

How then does one determine which effects are significant under these conditions? This is where the normal probability plot for effects becomes extremely useful. When effects are plotted on normal probability paper, the effects that are *not* significant will cluster around the value 0 in the middle with the significant ones separating out at the extremes. The significant effects are identified using a methodology developed by Lenth (1989)<sup>3</sup> (a discussion of which lies outside the intended scope of this chapter) and from there, an appropriate distribution fit can then be provided to the non-significant effects.

<sup>3</sup>Lenth, R.V. (1989). “Quick and Easy Analysis of Unreplicated Factorials,” *Technometrics*, 31, 469-473.



**FIGURE 19.8:** Normal probability plot for the effects, using Lenth's method to identify *A*, *D* and *AD* as significant.

Such a probability plot for this data set (using Lenth's method to determine the significance effects) is shown in Fig 19.8, where the effects *A* and *D*, respectively, Gap and Power, are identified as important, along with the two-way interaction *AD* (i.e., Gap\*Power).

At this point, it is important to pause and consider the repercussions of aliasing. According to the alias structure for this Resolution IV design, *AD* = *BC*, i.e., it is impossible from this data, to distinguish between the **Gap\*Power** interaction effect and the **Pressure\*Gas Flow** interaction effect. Thus, is the identified significant interaction the former (as indicated by default), or the latter? This is where “domain knowledge” becomes crucial in interpreting the results of fractional factorial data analysis. First, from pure common sense, if Gap and Power are identified as significant factors, it is far more natural and more likely that the two-way interaction that is also of significance will be **Gap\*Power** and not **Pressure\*Gas Flow**. This common sense conjecture is in fact corroborated by the physics of the process: the spacing between the anode and cathode (Gap) and the power applied to the cathode are far more likely to influence etch rate than the interaction between Pressure and Gas Flow rate, especially when none of these individual factors appear important by themselves. We therefore conclude, on the basis of the factorial analysis, aided especially by the normal probability plot of the effects that Gap and Power are the important factors that affect etch rate. This finding presents us with a fortuitous turn of events: we started with 4 potential factors, performed a set of 8 experiments based on a  $2^{4-1}$  fractional factorial design (with no replicates), and discovered that only 2 factors are significant. This immediately suggests projection. By projecting down onto the two relevant dimensions represented

by  $A$  and  $D$ , the 8 experiments will appear as if they were obtained from a  $2^2$  full factorial design (4 experiments) with 2 full sets of replicates, thereby allowing us to obtain estimates of the experimental error standard deviation, which in turn can be used to determine the precision of the effect estimates.

### Analysis Part II: Projection

As noted above, with only  $A$  and  $D$  as the surviving important factors, we can reanalyze the data only on these terms. Projection is carried out by removing all terms containing  $B$  and  $C$  including  $AB$   $AC$ , etc. from the "Terms" dialog box. Upon selecting the button labeled "Results" afterwards, one finds that the only surviving terms are the desired ones:  $A$ ,  $D$  and  $AD$ . When the analysis is repeated, we obtain the following results:

#### Factorial Fit: Etch Rate versus Gap, Power

##### Estimated Effects and Coefficients for Etch Rate (coded units)

| Term      | Effect | Coef    | SE Coef | T      | P            |
|-----------|--------|---------|---------|--------|--------------|
| Constant  |        | 756.00  | 7.494   | 100.88 | 0.000        |
| Gap       |        | -127.00 | -63.50  | 7.494  | -8.47 0.001  |
| Power     |        | 290.50  | 145.25  | 7.494  | 19.38 0.000  |
| Gap*Power |        | -197.50 | -98.75  | 7.494  | -13.18 0.000 |

S = 21.1955 PRESS = 7188

R-Sq = 99.36% R-Sq(pred) = 97.44% R-Sq(adj) = 98.88%

##### Analysis of Variance for Etch Rate (coded units)

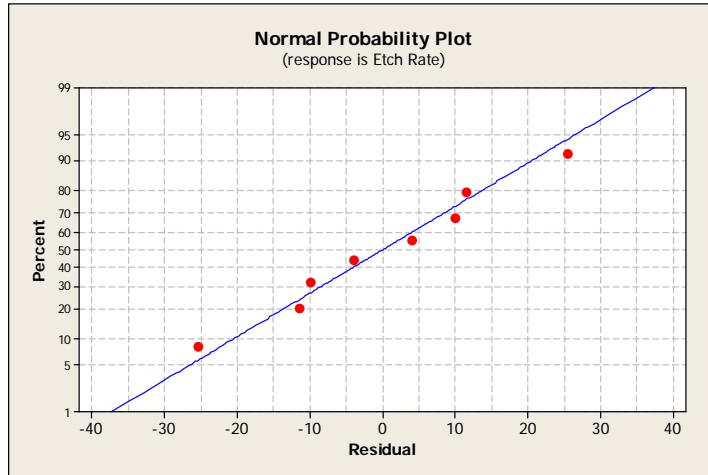
| Source             | DF | Seq SS | Adj SS | Adj MS | F      | P     |
|--------------------|----|--------|--------|--------|--------|-------|
| Main Effects       | 2  | 201039 | 201039 | 100519 | 223.75 | 0.000 |
| 2-Way Interactions | 1  | 78013  | 78013  | 78013  | 173.65 | 0.000 |
| Residual Error     | 4  | 1797   | 1797   | 449    |        |       |
| Total              | 7  | 280848 |        |        |        |       |

It should not be lost on the reader that the estimates have not changed; only now we have associated measures of their precisions. (The "SE Coef" term refers to the standard error associated with the coefficients, from which, as discussed in Chapters 14 and 15, one obtains 95% confidence intervals.) We also now have associated  $p$ -values. These additional quantities have been made available courtesy of the projection.

The conclusion is not only that the two most important factors are Gap and Power, but that the estimated relationship between Etch rate,  $y$ , and these factors, with  $x_1$  as Gap and  $x_2$  as Power, is :

$$y = 756.00 - 63.50x_1 + 145.25x_2 - 98.75x_1x_2 + \epsilon \quad (19.46)$$

The  $R^2$  value and other affiliated measures of the variability in the data explained by this simple model, are also contained in the MINITAB results shown above. These values indicate, among other things, that the amount of



**FIGURE 19.9:** Normal probability plot for the residuals of the Etch rate model in Eq (19.46) obtained upon projection of the experimental data to retain only the significant terms  $A$ , Gap ( $x_1$ ),  $D$ , Power ( $x_2$ ), and the interaction  $AD$ , Gap\*Power ( $x_1x_2$ ).

the variability in the data explained by this simple model is quite substantial. A normal probability plot of the estimated model residuals is shown in Fig 19.9, where visually, we see no reason to question the normality of the residuals.

## 19.8 Screening Designs: Plackett-Burman

As useful as fractional factorial designs are, the voids in the prescribed number of experimental runs widen substantially as the number of factors  $k$  increases. This is because the experimental runs are limited to powers of 2 (i.e., 4, 8, 16, 32, 64, 128 etc.). This leaves the experimenter with limited options for larger  $k$ . For example, for  $k = 7$ , if the total number of runs,  $N = 128$  are too many, the next options are 64, or 32.

If necessity is the mother of invention, then it is no surprise that these voids were filled in 1946 by two British scientists, Robin L. Plackett and J. Peter Burman<sup>4</sup>, who were part of a team working on the development of anti-aircraft shells during the World War II bombing of London. The time constraint faced by the team necessitated the screening of a very large number of potential factors very quickly. The result is that voids left by fractional

<sup>4</sup>Plackett, R.L. and J. P. Burman (1946). “The design of optimum multifactorial experiments.” *Biometrika*, 33, 305–325

factorial designs can now be filled by what has become known appropriately as Plackett-Burman (PB) designs because they involve  $2r$  experimental runs, where fractional factorial designs involve  $2^r$  experimental runs. Thus, with PB designs, experimental runs of sizes  $N = 12, 20, 24, 28$ , etc. are now possible.

### 19.8.1 Primary Characteristics

PB designs are also 2-level designs (like fractional factorials), but where fractional factorial designs involve runs that are *powers* of 2, PB designs have experimental runs that are multiples of 2. All PB designs are of Resolution III, however, so that *all* main effects are aliased with two-way interactions. They should not be used therefore when 2-way interactions might be as important as main effects.

PB designs are best used to screen for critical main effects when the number of potential factors is large. The primary advantage is that they involve remarkably few runs for a large number of factors; they are therefore extremely efficient and cost-effective. For example, the following is the design table for a PB design of 12 runs for  $k = 7$  factors!

| Run # | A | B | C | D | E | F | G |
|-------|---|---|---|---|---|---|---|
| 1     | + | - | + | - | - | - | + |
| 2     | + | + | - | + | - | - | - |
| 3     | - | + | + | - | + | - | - |
| 4     | + | - | + | + | - | + | - |
| 5     | + | + | - | + | + | - | + |
| 6     | + | + | + | - | + | + | - |
| 7     | - | + | + | + | - | + | + |
| 8     | - | - | + | + | + | - | + |
| 9     | - | - | - | + | + | + | - |
| 10    | + | - | - | - | + | + | + |
| 11    | - | + | - | - | - | + | + |
| 12    | - | - | - | - | - | - | - |

The main disadvantages have to do with their resolution: with PB designs, it is difficult, if not entirely impossible, to determine interaction effects independently. Furthermore the alias structures are quite complicated (but important). The designs have also been known to be prone to poor precision, but this can be mitigated with the use of replicates. It is recommended that computer software be used for both design and analysis of experiments using the PB strategy.

### 19.8.2 Design and Analysis

Programs such as MINITAB will generate PB designs and the accompanying alias structure. Because they are orthogonal two-level designs, the analysis

is similar to that for factorials. But, we emphasize again, that these are best carried out using computer programs.

Additional discussions are available in Chapter 7 of the Box, Hunter and Hunter, (2005) reference provided earlier. An application of PB designs in biotechnology discussed in Balusu, *et al.*, 2004<sup>5</sup> is highly recommended to the interested reader.

---

## 19.9 Response Surface Designs

Frequently, the objective of the experimental study is to capture the relationship between the response  $y$  and the factors  $x_i$  mathematically so that the resulting model can be used to optimize the response with respect to the factors. Under these circumstances, for such models to be useful, they will have to include more than the approximate linear ones possible with two-level designs. Response surface methodology is the approach for obtaining models of the sort required for optimization studies. They provide the designs for efficiently fitting more complex models to represent the relationship between the response,  $Y$ , and the factors, with the resulting model known as the “response surface.” A detailed discussion is impossible in this lone section of an introductory chapter devoted to the topic. The classic reference is the book by Box and Draper<sup>6</sup> that the interested reader is encouraged to consult. What follows is a summary of the salient features of this experimental design strategy that finds important applications in engineering practice.

### 19.9.1 Characteristics

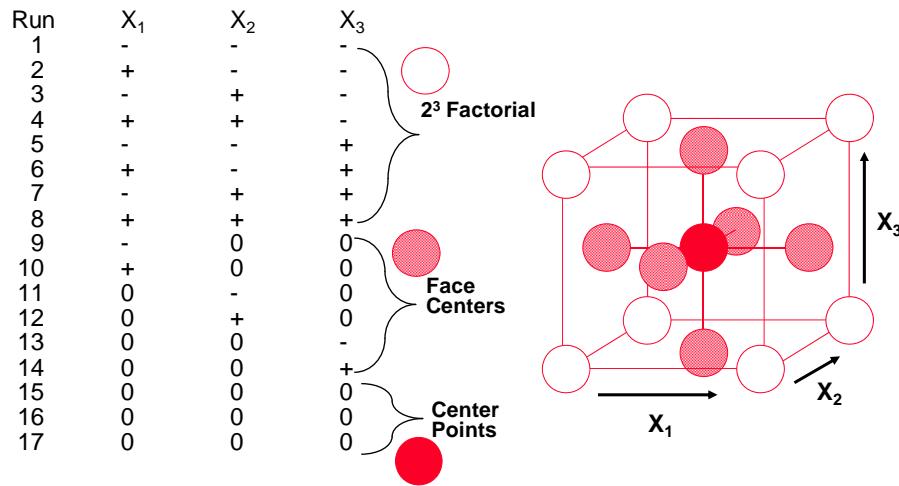
With response surface designs, each factor is evaluated at three settings, generically denoted “−”, “0” and “+”, the Low, Medium and High levels. These designs are most applicable when factors are continuous variables as opposed to categorical variables because the response surfaces are assumed to be smooth. This is not the case with categorical variables; for example, there is no “interval” to speak of between Type A and Type B fertilizer, as opposed to the interval between temperature at 100°C and 250°C, which is continuous.

The key assumption is that the relationship between response and factors (i.e., the response surface) can be approximated well by low-order polynomials; often, no more than second-order polynomials are used to capture any response

---

<sup>5</sup>Balusu R., R. M. R. Paduru, G. Seenayya, and G. Reddy, (2004). Production of ethanol from *Clostridium thermocellum* SS19 in submerged fermentation: Screening of nutrients using Plackett-Burman design. *Applied Biochem. & Biotech.*, **117** (3) 133-142.

<sup>6</sup>G.E.P. Box and N.R. Draper (1987). *Empirical Model Building and Response Surfaces*, J. Wiley, N.Y.



**FIGURE 19.10:** The 3-factor face-centered cube (FCC) response surface design and its constituent parts:  $2^3$  factorial base, Open circles; face center points, lighter shaded circles; center point, darker solid circle.

surface curvature. For example, the typical postulated model for two-factors is

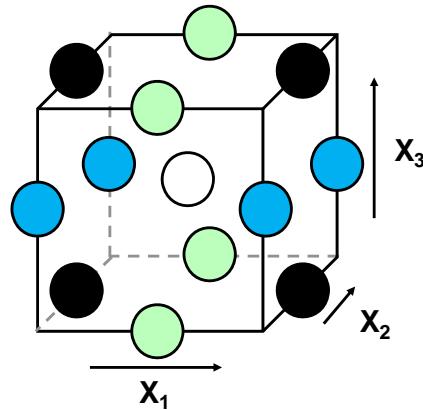
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \epsilon \quad (19.47)$$

### 19.9.2 Response Surface Designs

Of the many available response surface designs, the face-centered cube (FCC) design (part of the family of Central Composite designs) and the Box-Behnken design are the most commonly employed. The face-centered cube design, as its name implies, is based upon adding axial and central experimental points to the basic  $2^k$  factorial design. For example, for 3 factors, the face-centered cube design is shown in Fig 19.10. The three components of the design are represented with circles: the open circles represent the base  $2^3$  factorial design; the lighter shaded circles are the face centers, with the darker solid circle as the dead center of the cube. The standard design shown calls for replicating the center point three times for a total of 17 runs.

When corner points are infeasible, (for example, when the high temperature, high pressure and high concentration combination will lead to an explosion, and/or when the low temperature, low pressure, and low catalyst concentration will lead to reaction extinction) the Box-Behnken design is a viable option. Because this design is based not on face centers but on edge-centers, it avoids the potential problems caused by infeasible corner points. An example 3-factor, Box-Behnken design is shown in Fig 19.11. The design is based on three  $2^2$  factorial designs moved to the edge center of the third

| Run | $X_1$ | $X_2$ | $X_3$ |  |
|-----|-------|-------|-------|--|
| 1   | -     | -     | 0     |  |
| 2   | +     | -     | 0     |  |
| 3   | -     | +     | 0     |  |
| 4   | +     | +     | 0     |  |
| 5   | -     | 0     | -     |  |
| 6   | +     | 0     | -     |  |
| 7   | -     | 0     | +     |  |
| 8   | +     | 0     | +     |  |
| 9   | 0     | -     | -     |  |
| 10  | 0     | +     | -     |  |
| 11  | 0     | -     | +     |  |
| 12  | 0     | +     | +     |  |
| 13  | 0     | 0     | 0     |  |
| 14  | 0     | 0     | 0     |  |
| 15  | 0     | 0     | 0     |  |



**FIGURE 19.11:** The 3-factor Box-Behnken response surface design and its constituent parts:  $X_1, X_2$ :  $2^2$  factorial points moved to the center of  $X_3$  to give the darker shaded circles at the edge-centers of the  $X_3$  axes;  $X_2, X_3$ :  $2^2$  factorial points moved to the center of  $X_1$  to give the lighter shaded circles at the edge-centers of the  $X_1$  axes;  $X_1, X_3$ :  $2^2$  factorial points moved to the center of  $X_2$  to give the solid circles at the edge-centers of the  $X_2$  axes; the center point, open circle.

factor and point at the dead center of the cube. Thus, the first four runs are based on the four  $X_1, X_2$  factorial points moved to the center of  $X_3$  to give the dark shaded circles at the edge-centers of the  $X_3$  axes. The next four are based on the four four  $X_2, X_3$  factorial points moved to the center of  $X_1$  to give the lighter shaded circles at the edge-centers of the  $X_1$  axes. The next four likewise are based on the to give the  $X_1, X_3$  factorial points moved to the center of  $X_2$  solid circles at the edge-centers of the  $X_2$  axes. The center points are the open circles. The design also calls for three replicates of the center points. Note that while the three-factor FCC design involves 17 points, the Box-Behnken design involves 15.

### 19.9.3 Design and Analysis

As with every experimental design discussed in this chapter, programs such as MINITAB will generate response surface designs and also carry out the necessary analysis once the data has been collected. It is important to note that in coded form (using  $-1, +1$  and 0), the design matrix is orthogonal. The response surface analysis is therefore similar to polynomial regression discussed in Chapter 16, but it is especially simplified because the design matrix is orthogonal.

Additional discussions are available in the Box and Draper reference given earlier, and also, for example in Chapter 12 of Box, Hunter, and Hunter,

2005, and Chapter 10 of Ryan, 2007<sup>7</sup>. An application to catalyst design may be found in Hendershot, *et al.*, 2004<sup>8</sup>. Another application in food engineering is available in Pericin, D. *et al.*, 2007<sup>9</sup>. An application to industrial process optimization is discussed as a case study in Chapter 20.

## 19.10 Introduction to Optimal Designs

### 19.10.1 Background

Whether stated explicitly or not, all statistical experimental design strategies are based on assumed mathematical models. To buttress this point, at every stage of the discussion in this chapter, we have endeavored to state explicitly the postulated model underlying each design. Most of the models we have encountered have been relatively simple, involving at most polynomials of modest order. What happens if an experimenter has available a model for the system under investigation, and the objective is to estimate the unknown model parameters using experimental data? Under such circumstances, the appropriate experimental design should be one that provides — for the given model — the “best” set of experimental conditions so that the acquired data is “most informative” for the task at hand: estimating the unknown model parameters. This is the motivation behind optimal designs.

If the postulated model is of the form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon} \quad (19.48)$$

i.e., the matrix form of the general linear regression model, we may recall that, given the data vector,  $\mathbf{y}$ , and the “design matrix”  $\mathbf{X}$ , the least-squares estimate of the parameter vector is:

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (19.49)$$

Optimal experimental designs for problems of this kind are concerned with determining values for the elements of the matrix  $\mathbf{X}$  that will provide estimates in Eq (19.49) that are “optimal” in some specific sense. Not surprisingly, the optimality criteria are usually related to the matrix

$$\mathcal{F}_I = (\mathbf{X}^T \mathbf{X}) \quad (19.50)$$

known as the Fisher information matrix.

<sup>7</sup>Ryan, T.P. (2007). *Modern Experimental Design*, John Wiley, New Jersey.

<sup>8</sup>R. J. Hendershot, W. B. Rogers, C. M. Snively, B. A. Ogunnaike, and J. Lauterbach, (2004). “Development and optimization of NOx storage and reduction catalysts using statistically guided high-throughput experimentation, *Catalysis Today* 98, 375385.

<sup>9</sup>Pericin, D. *et al.*, (2007). “Evaluation of solubility of pumpkin seed globulins by response surface method.” *J. Food Engineering*. 84, 591-594.

### 19.10.2 “Alphabetic” Optimal Designs

The “D-Optimal” design selects values of the factors  $x$  to maximize  $|\mathbf{X}^T \mathbf{X}|$ , the determinant of the information matrix. This optimization criterion maximizes the information content of the data and hence of the estimate. It is possible to show that for the factorial models in Eq (19.36), with  $k$  factors, the “D-Optimal” design is precisely the  $2^k$  factorial design, where  $-1$  and  $+1$  represent, respectively, the left and right extremes of the feasible region for each factor, giving a  $k$ -dimensional hypercube in the experimental space.

Other optimality criteria give rise to variations in the optimal design arsenal. Some of these are listed below, (in alphabetical order!), the optimality criteria themselves and what they mean for the parameter estimates:

1. **A-Optimal designs:** minimize the trace of  $(\mathbf{X}^T \mathbf{X})^{-1}$ . And because (as we may recall from Chapter 16),

$$Var(\hat{\boldsymbol{\theta}}) = (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2 \quad (19.51)$$

with  $\sigma^2$  as the error variance, then the “A-optimality” criterion minimizes the average variance of the estimated coefficients, a desirable property.

Since the inverse of a matrix is its adjoint divided by its determinant, it should be clear therefore that the “D-Optimality” criterion minimizes the general variance of the estimated parameters.

2. **E-Optimal designs:** maximize the smallest eigenvalue of  $(\mathbf{X}^T \mathbf{X})$ . This is essentially the same as maximizing the condition number of the design matrix  $\mathbf{X}$ , preventing the sort of ill-conditioning that leads to poor estimation. This is a less-known, and less popular design criterion.
3. **G-Optimal designs:** minimize the maximum diagonal element of the hat matrix defined in Chapter 16 as  $\mathbf{H} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ , which, as we may recall, is associated with the model’s predicted values. The “G-optimality” criterion minimizes the maximum variance associated with the predicted values,  $\hat{\mathbf{y}}$ .
4. **V-Optimal designs:** minimize the trace (sum of the diagonal elements) of the hat matrix. As such, the “V-optimality” criterion minimizes the average prediction variance associated with  $\hat{\mathbf{y}}$ .

Even with such a simple linear model, the computations involved in obtaining these optimal designs are not trivial; and things become even more complicated when the models are non-linear in the parameters. Also, these optimal designs have come under some criticism for various reasons (see Chapter 13 of Ryan, (2007)). Nevertheless, they have been put to good use in some chemical

engineering applications, for example, Asprey and Macchietto, (2000)<sup>10</sup> and (2002)<sup>11</sup>. They have also inspired experimental design techniques for such challenging problems as signal transduction model development in molecular biology (Birtwistle, *et al.*, (2009)<sup>12</sup>). The most current definitive text on the subject matter is Atkinson *et al.*, (2007)<sup>13</sup>, which is recommended highly to the interested reader.

---

### 19.11 Summary and Conclusions

This chapter has been primarily concerned with providing something of an extended overview of strategies for designing experiments that will generate informative data sets most efficiently. The discussions took us rapidly from designs for simple, one-factor experiments through the popular factorial and fractional factorial designs. And even though the coverage is by no means comprehensive in terms of breadth, there is sufficient detail in what was presented. The discussions about sample size issues and about folding and projection may have been brief but they are important, with significant implications in practice. Some in-chapter examples and a handful of end-of-chapter exercises and applications problems illustrate this point. The deliberately abbreviated discussions of Plackett-Burman designs and response surface methods should be viewed as an attempt to whet the appetite of the interested reader; and if such a reader then chooses to pursue the topic further (for example, in textbooks dedicated to these matters), the introduction in this chapter should facilitate such an endeavor.

With the discussion of experimental designs in this chapter complementing the earlier discussions of both descriptive statistics (Chapter 12) and inductive (i.e., inferential) statistics (Chapters 13–18), our intended coverage of the very broad subject matter of statistics is now formally complete. Even though we have presented principles and illustrated mechanics and applications with appropriate examples at various stages of our coverage, an appropriate capstone that will serve to consolidate all the material is still desirable. The next chapter (Chapter 20) provides such a capstone. It contains a series of case studies carefully chosen to demonstrate the application of the various aspects

<sup>10</sup>Asprey, S. and Macchietto, S. (2000). "Statistical tools for optimal dynamic model building." *Computers and Chemical Engineering* 24, 1261-1267.

<sup>11</sup>Asprey, S. and Macchietto, S.(2002). "Designing robust optimal dynamic experiments." *Journal of Process Control* 12, 545-556.

<sup>12</sup>Birtwistle, M. R., B. N. Kholodenko, and B. A. Ogunnaike, (2009). "Experimental Design for Parameter Identifiability in Biological Signal Transduction Modeling," Chapter 10, *Systems Analysis of Biological Networks*, Ed. A. Jayaraman and J. Hahn, Artech House, London.

<sup>13</sup>Atkinson, A. A. Donev, and R. Tobias, (2007). *Optimum Experimental Designs, with SAS*, Oxford University Press, NY.

of statistics—graphical analysis; estimation; hypothesis testing; regression; experimental design—in real-life applications.

---

## REVIEW QUESTIONS

- 1.** In what sense is the term “experimental studies” used in this chapter?
- 2.** What is an observational study?
- 3.** What is the key distinguishing characteristic of the experimental studies of concern in this chapter?
- 4.** What are the two basic tasks involved in the experimental studies discussed in this chapter? What complicates the effective execution of these tasks?
- 5.** How does “statistical design of experiments” enable efficient conduct of experimental studies?
- 6.** List the phases of efficient experimental studies and what each phase entails.
- 7.** In the terminology of statistical design of experiments, what is a *response*, a *factor*, a *level*, and a *treatment*?
- 8.** When more than two population means are to be compared simultaneously, why are multiple pairwise *t*-tests not recommended?
- 9.** What is ANOVA, and on what is it predicated?
- 10.** What are the two central assumptions in ANOVA?
- 11.** What is a one-way classification of a single factor experiment, and how is it different from a two-way classification?
- 12.** What is the postulated model and what are the hypotheses for a one-way classification experiment?
- 13.** What is the “completely randomized” design, and why is it appropriate for the one-way classification experiment?
- 14.** What makes a one-way classification experimental design balanced as opposed to unbalanced?
- 15.** What is the ANOVA identity, and what is its primary implication in data analysis?

- 16.** What is the difference between a “fixed effect” and a “random effect” ANOVA?
- 17.** What is the Kruskal-Wallis test?
- 18.** What is the randomized complete block design?
- 19.** What is the postulated model and what are the hypotheses for the randomized complete block design?
- 20.** What is a “nuisance” variable?
- 21.** What is the difference between a 2-way classification of a single factor experiment and a two-factor experiment?
- 22.** What is the postulated model and what are the hypotheses for a two-factor experiment?
- 23.** What is the potential problem with general multi-level multi-factor experiments?
- 24.** What is a  $2^k$  factorial experiment?
- 25.** What are some of the advantages and disadvantages of  $2^k$  factorial designs?
- 26.** What does it mean that  $2^k$  factorial designs are balanced *and* orthogonal?
- 27.** What is the general procedure for carrying out  $2^k$  factorial experiments?
- 28.**  $2^k$  factorial designs are best applied to what kinds of problems?
- 29.** What are screening designs and what is the rationale behind them?
- 30.** In fractional factorial designs, what is “aliasing,” and what is an “alias structure”?
- 31.** What is a defining relation?
- 32.** What is the resolution of a fractional factorial design?
- 33.** What are the main characteristics of a Resolution III, a Resolution IV, and a Resolution V design?
- 34.** If one is interested in estimating both main effects and 2-way interactions, why should one not use a Resolution III design?
- 35.** What is “projection,” and under what condition is it possible?
- 36.** What is “folding”?

- 37.** What problem with fractional factorial design is ameliorated with Plackett-Burman designs?
- 38.** What is the resolution of all Plackett-Burman designs?
- 39.** What are Plackett-Burman designs best used for?
- 40.** What are response surface designs and what distinguishes them from basic  $2^k$  factorial designs?
- 41.** What is a typical response surface design model for a two-factor experiment?
- 42.** What is the difference between a face centered cube design and a Box-Behnken design?
- 43.** When is a Box-Behnken design to be preferred over a face centered cube design?
- 44.** What is an “optimal” experimental design?
- 45.** What is the Fisher information matrix for a linear regression model?
- 46.** What optimization criteria lead respectively to D-Optimal, A-Optimal, E-Optimal, G-Optimal and V-Optimal designs?

---

## EXERCISES

**19.1** In each of the following, identify the response, the factors, the levels, and the total number of treatments. Also identify which variables are categorical and which are quantitative.

- (i) A chemical engineering catalysis researcher is interested in the effect of  $NO$  concentration, (3500 ppm, 8650 ppm);  $O_2$  concentration (4%, 8%); CO concentration, (3.5%, 5.5%); Space velocity, (30,000, 42,500) mL/hr/ $g_{cat}$ ; Temperature, (548 K, 648 K);  $SO_2$  concentration, (0 ppm, 300ppm); and Catalyst metal type, (Pt, Ba, Fe), on saturation  $NO_x$  storage ( $\mu$  mol).
- (ii) A material scientist studying a reactive extrusion process is interested in the effect of screw speed, (135 rpm, 150 rpm), feed-rate, (15 lb/hr, 25 lb/hr), and feed-composition, %A, (25, 30, 45) on the residence time distribution,  $f(t)$ .
- (iii) A management consultant is interested in the risk-taking propensity of three types of managers: entrepreneurs, newly-hired managers and newly promoted managers.
- (iv) A child psychologist is interested in the effect of socio-economic status of parents (Lower class, Middle class, Upper class), Family size (Small, Large) and Mother’s marital status (Single-never married, Married, Divorced), on the IQ of 5-year-olds.

**19.2** Consider the postulated model for the single-factor, completely randomized experiment given in Eq (19.2):

$$Y_{ij} = \mu_j + \epsilon_{ij}; i = 1, 2, \dots, n_j; j = 1, 2, \dots, k$$

where  $\mu_j$  is the mean associated with the  $j^{th}$  treatment; furthermore, let the grand mean of the complete data set be  $\mu$ .

(i) If the  $j^{th}$  treatment mean is expressed as:

$$\mu_j = \mu + \tau_j; j = 1, 2, \dots, k$$

so that  $\tau_j$  represents the  $j^{th}$  treatment effect, show that

$$\sum_{j=1}^k \tau_j = 0 \quad (19.52)$$

and hence establish Eq (19.52).

(ii) With the randomized complete block design, the postulated model was given as:

$$Y_{ij} = \mu + \tau_j + \beta_i + \epsilon_{ij}; i = 1, 2, \dots, r; j = 1, 2, \dots, k$$

where  $\tau_j$ , is the  $j^{th}$  treatment effect;  $\beta_i$  is the  $i^{th}$  block effect and  $\epsilon_{ij}$  is the random error. Show that

$$\sum_{j=1}^k \tau_j = 0; \sum_{i=1}^r \beta_i = 0$$

and hence establish Eq (19.20).

**19.3** To demonstrate the mechanics of the ANOVA decomposition, consider the following table of obviously “made-up” data for a fictitious completely randomized design experiment (one-way classification) involving 2 levels of a single factor:

| Sample | Level 1 | Level 2 |
|--------|---------|---------|
| 1      | 1       | 4       |
| 2      | 2       | 5       |
| 3      | 3       | 6       |

(i) obtain, as illustrated in the text, the following (6-dimensional) vectors:

$\mathbf{E}_Y$ , whose elements are  $(Y_{ij} - \bar{Y}_{..})$ ;

$\mathbf{E}_T$ , whose elements are  $(Y_{..j} - \bar{Y}_{..})$ ; and

$\mathbf{E}_E$ , whose elements are  $(Y_{ij} - \bar{Y}_{..j})$ .

(ii) Confirm the error decomposition identity:

$$\mathbf{E}_Y = \mathbf{E}_T + \mathbf{E}_E \quad (19.53)$$

(iii) Confirm the orthogonality of  $\mathbf{E}_T$  and  $\mathbf{E}_E$ , i.e. that,

$$\mathbf{E}_T^T \mathbf{E}_E = 0 \quad (19.54)$$

(iv) Finally confirm the “sum-of-squares” identity:

$$\|\mathbf{E}_Y\|^2 = \|\mathbf{E}_T\|^2 + \|\mathbf{E}_E\|^2 \quad (19.55)$$

where the squared norm of an  $n$ -dimensional vector,  $\mathbf{a}$ , whose elements are  $a_i : i = 1, 2, \dots, n$ , is defined as:

$$\|\mathbf{a}\|^2 = \sum_{i=1}^n a_i^2 \quad (19.56)$$

**19.4** Consider the response  $Y_{ij}$  in a single-factor completely randomized experiment. From the definition given in the text of the treatment average,  $\bar{Y}_{.j}$ , and the grand average,  $\bar{Y}_{..}$ , the “error decomposition” expression given in Eq (19.11) is

$$\begin{aligned} (Y_{ij} - \bar{Y}_{..}) &= (\bar{Y}_{.j} - \bar{Y}_{..}) + (Y_{ij} - \bar{Y}_{.j}) \\ E_Y &= E_T + E_E \end{aligned}$$

Take sums-of-squares in this equation and show that the result is the following sums of squares identity:

$$\begin{aligned} \sum_{j=1}^k \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{..})^2 &= N \sum_{j=1}^k (\bar{Y}_{.j} - \bar{Y}_{..})^2 + \sum_{j=1}^k \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{.j})^2 \\ SS_Y &= SS_T + SS_E \end{aligned}$$

and hence establish Eq (19.12).

**19.5** Consider the data shown in the table below, where  $F_i : i = 1, 2, 3, 4$  represents four factors; and  $B_i : i = 1, 2, 3, 4, 5$  represents 5 block. The data were generated from normal populations with mean 10 and standard deviation 2, except for column 3 for  $F_3$ , generated from a normal distribution with mean 12 and standard deviation 2. Also, the first row was adjusted by adding 1.5 across the entire row.

|       | $F_1$ | $F_2$ | $F_3$ | $F_4$ |
|-------|-------|-------|-------|-------|
| $B_1$ | 14.5  | 11.5  | 12.7  | 10.9  |
| $B_2$ | 10.6  | 10.2  | 13.6  | 10.8  |
| $B_3$ | 12.0  | 9.1   | 14.6  | 9.7   |
| $B_4$ | 9.0   | 8.8   | 12.2  | 8.7   |
| $B_5$ | 11.6  | 10.6  | 15.3  | 10.0  |

- (i) First analyze the data as a one-way classification with five replicates and comment on the results, especially the  $p$ -value associated with the ANOVA  $F$ -test. Note the  $R^2$  and  $R^2_{adj}$  values. What do these values indicate about how much of the variation in the data has been explained by the one-way classification model?
- (ii) Now analyze the data as a two-way classification, where the blocks are now explicitly recognized as such. Compare the results with those obtained in (i) above. Comment on what this exercise indicates about what can happen when a nuisance effect is not explicitly separated out from an analysis of a one-factor experiment.

**19.6** Refer to Exercise 19.5 and the supplied data. Repeat the analysis, this time saving the residuals in each case (one-way first, and then two-way next). Carry out a normality test on both sets of residuals, plot both residuals on the same graph, and compare their standard deviations. Comment on what these residuals imply about which ANOVA model more appropriately fits the data, especially in light of what is known about how this particular data set was generated.

**19.7** Refer to Example 19.2 in the text. Repeat the data analysis in the example and save the residuals for analysis. Assess the normality of the residuals. Interpret the results and discuss what they imply about the ANOVA model for the Tire wear data.

**19.8** Write out the model for a  $2^4$  factorial design where the factors are  $x_1, x_2, x_3$  and  $x_4$  and the response is  $y$ .

(i) How many parameters are to be estimated in order to specify this model completely?

(ii) Obtain a base design for this experiment.

**19.9** For each of the following base factorial designs, specify the number of replicates required.

(i)  $2^3$  with signal-to-noise ratio specified as  $\rho_{SN} = 1.5$

(ii)  $2^2$  with signal-to-noise ratio specified as  $\rho_{SN} = 1.5$

(iii)  $2^2$  with signal-to-noise ratio specified as  $\rho_{SN} = 2.0$

(iv)  $2^4$  with signal-to-noise ratio specified as  $\rho_{SN} = 1.5$

**19.10** A factorial experiment is to be designed to study the effect on reaction yield of temperature  $x_1$  at  $180^\circ C$  and  $240^\circ C$  in conjunction with Pressure at 1 atm and 2 atm. Obtain a design in terms of the original variables with three full replicates.

**19.11** The design matrix for a  $2^2$  factorial design for factors  $X_1$  and  $X_2$  is shown below along with the measured responses  $y_i : i = 1 - 4$ .

| Run | $X_1$ | $X_2$ | Response |
|-----|-------|-------|----------|
| 1   | -1    | -1    | $y_1$    |
| 2   | 1     | -1    | $y_2$    |
| 3   | -1    | 1     | $y_3$    |
| 4   | 1     | 1     | $y_4$    |

Given model equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$$

it is possible to use the regression approach of Chapter 16 to obtain the estimates of the model parameters as follows:

(i) Write the model equation in vector-matrix form

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \mathbf{X} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_{12} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{bmatrix}$$

i.e., as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon} \quad (19.57)$$

From the design table given above, determine the matrix  $\mathbf{X}$  in terms of -1 and 1 .

(ii) From Chapter 16, we know that the least squares estimate of the unknown parameters in Eq (19.57) is:

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

Show that for this particular  $2^2$  factorial design model, the least squares solution,  $\hat{\theta}$ , is given as follows:

$$\begin{aligned}\hat{\beta}_0 &= \frac{1}{4}(y_1 + y_2 + y_3 + y_4) \\ \hat{\beta}_1 &= \frac{1}{4}(-y_1 + y_2 - y_3 + y_4) \\ \hat{\beta}_2 &= \frac{1}{4}(-y_1 - y_2 + y_3 + y_4) \\ \hat{\beta}_{12} &= \frac{1}{4}(y_1 - y_2 - y_3 + y_4)\end{aligned}$$

- (iii) Examine the design table shown above and explicitly associate the elements of this design table directly with the least squares solution in (ii); identify how such a solution can be obtained directly from the table without necessarily going through the least squares computation.

**19.12** The following data was obtained from a  $2^2$  factorial experiment for factors  $A$  and  $B$  with two replicate runs. Analyze the data and estimate the main effects and the interaction term.

| Run |     |     | Response    |             |
|-----|-----|-----|-------------|-------------|
|     | $A$ | $B$ | Replicate 1 | Replicate 2 |
| 1   | -1  | -1  | 0.68        | 0.65        |
| 2   | 1   | -1  | 3.81        | 4.03        |
| 3   | -1  | 1   | 1.67        | 1.71        |
| 4   | 1   | 1   | 8.90        | 9.66        |

**19.13** Generate a design table for a  $2^4$  factorial experiment for factors  $A, B, C$  and  $D$ , to be used as a base design for a  $2^{5-1}$  half factorial design which now includes a fifth factor  $E$ .

- (i) Use  $ABC = E$  to generate the  $2^{5-1}$  design and show the resulting table. Obtain the alias structure for the remaining 3-factor interactions. What is the resolution of this design? Which main effect can be estimated with no confounding?  
(ii) This time, use  $BCD = E$  to generate the  $2^{5-1}$  design and repeat (i).

**19.14** The following table shows the result of a full 32-run,  $2^5$  factorial experiment involving 5 factors  $A, B, C, D$  and  $E$ .

| Run | <i>A</i> | <i>B</i> | <i>C</i> | <i>D</i> | <i>E</i> | <i>y</i> |
|-----|----------|----------|----------|----------|----------|----------|
| 1   | -1       | -1       | -1       | -1       | -1       | 2.98     |
| 2   | 1        | -1       | -1       | -1       | -1       | 0.05     |
| 3   | -1       | 1        | -1       | -1       | -1       | 3.30     |
| 4   | 1        | 1        | -1       | -1       | -1       | 7.92     |
| 5   | -1       | -1       | 1        | -1       | -1       | -0.75    |
| 6   | 1        | -1       | 1        | -1       | -1       | 6.58     |
| 7   | -1       | 1        | 1        | -1       | -1       | -1.08    |
| 8   | 1        | 1        | 1        | -1       | -1       | 15.48    |
| 9   | -1       | -1       | -1       | 1        | -1       | 2.64     |
| 10  | 1        | -1       | -1       | 1        | -1       | -0.90    |
| 11  | -1       | 1        | -1       | 1        | -1       | 3.37     |
| 12  | 1        | 1        | -1       | 1        | -1       | 6.72     |
| 13  | -1       | -1       | 1        | 1        | -1       | -1.05    |
| 14  | 1        | -1       | 1        | 1        | -1       | 7.18     |
| 15  | -1       | 1        | 1        | 1        | -1       | -0.97    |
| 16  | 1        | 1        | 1        | 1        | -1       | 14.59    |
| 17  | -1       | -1       | -1       | -1       | 1        | 3.14     |
| 18  | 1        | -1       | -1       | -1       | 1        | -1.09    |
| 19  | -1       | 1        | -1       | -1       | 1        | 3.11     |
| 20  | 1        | 1        | -1       | -1       | 1        | 7.37     |
| 21  | -1       | -1       | 1        | -1       | 1        | -1.32    |
| 22  | 1        | -1       | 1        | -1       | 1        | 6.53     |
| 23  | -1       | 1        | 1        | -1       | 1        | -0.60    |
| 24  | 1        | 1        | 1        | -1       | 1        | 14.25    |
| 25  | -1       | -1       | -1       | 1        | 1        | 2.93     |
| 26  | 1        | -1       | -1       | 1        | 1        | -0.51    |
| 27  | -1       | 1        | -1       | 1        | 1        | 3.46     |
| 28  | 1        | 1        | -1       | 1        | 1        | 6.69     |
| 29  | -1       | -1       | 1        | 1        | 1        | -1.35    |
| 30  | 1        | -1       | 1        | 1        | 1        | 6.59     |
| 31  | -1       | 1        | 1        | 1        | 1        | -0.82    |
| 32  | 1        | 1        | 1        | 1        | 1        | 15.53    |

- (i) Estimate all the main effects and interactions. (Of course, you should use a computer program.)
- (ii) Since no replicates are provided, use the normal probability plot and Lenth's method (which should be available in your computer program) to confirm that only the main effects, *A*, *B*, and *C*, and the two-way interactions, *AB* and *AC*, are significant.
- (iii) In light of (ii) "project" the data down appropriately and reanalyze the data. This time note the uncertainty estimates associated with the estimates of the main effects and interactions; note also the various associated  $R^2$  values and comment on the fit of the reduced model to the data.

**19.15** Refer to Exercise 19.14 and the accompanying data table. This time create a  $2^{5-1}$  fractional factorial design and pretend that the experimenter had used this fractional factorial design instead of the full one. Extract from the full 32-run data table the results corresponding to the 16 runs indicated in the created  $2^{5-1}$  design.

- (i) Determine the design resolution and the alias structure.
- (ii) Repeat the entire Exercise 19.14 for *only* the indicated 16 experimental re-

sults. Compare the results of the analysis with that in Exercise 19.14. Does the experimenter lose anything significant by running only half of the full  $2^5$  factorial experiments? Discuss what this particular example illustrates regarding the use of fractional factorial designs when the experimental study involves many factors.

**19.16** Refer to Exercise 19.14 and the accompanying data table. Create an 8-run,  $2^{5-2}$  fractional factorial design and, again as in Exercise 19.15, pretend that the experimenter has used this design instead of the full one. Extract from the full 32-run data table, the results corresponding to the 8 runs indicated in the  $2^{5-2}$  design.

- Determine the design resolution and alias structure. Can any two-way interactions be determined independently without confounding? What does this imply in terms of what can be truly estimated from this much reduced set of results?
- Repeat the entire Exercise 19.14. How do the estimates of the main effects compare to the ones obtained from the full data set?
- If the experimenter had only been interested in determining which of the 5 factors has a significant effect on the response, comment on the advantages/disadvantages of the quarter fraction, 8-run design versus the full 32-run experimental design.

**19.17** Consider an experiment to determine the effect of two factors,  $x_1$  and  $x_2$ , on the response  $Y$ . Further, consider that the postulated model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$$

Let the minimum allowable values for each factors be represented in coded form as  $-1$  and the maximum allowable values as  $1$ . In this case, the  $2^2$  factorial design recommends the following settings for the factors:

| $X_1$ | $X_2$ |
|-------|-------|
| -1    | -1    |
| 1     | -1    |
| -1    | 1     |
| 1     | 1     |

- Show that for this  $2^2$  design, if the model is written in the matrix form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \epsilon$$

the Fisher information matrix (FIM) is given by

$$\mathcal{F}_I = (\mathbf{X}^T \mathbf{X}) = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

a diagonal matrix; hence establish that the determinant is given by:

$$|\mathbf{X}^T \mathbf{X}| = 4^4$$

- For any other selection of orthogonal experimental points for  $x_1$  and  $x_2$ , such as shown in the table below,

| $X_1$     | $X_2$     |
|-----------|-----------|
| $-\alpha$ | $-\alpha$ |
| $\alpha$  | $-\alpha$ |
| $-\alpha$ | $\alpha$  |
| $\alpha$  | $\alpha$  |

where  $0 < \alpha < 1$ , show that the determinant of the FIM will be

$$|\mathbf{X}^T \mathbf{X}| = (4\alpha^2)^4$$

Because  $0 < \alpha < 1$ , this determinant will be significantly less than the determinant of the FIM for the factorial design. (Since it can be shown that any other selection of four non-orthogonal points in the region bounded by the rectangle in  $-\alpha < x_1 < \alpha; -\alpha < x_2 < \alpha$  in the  $x_1$ - $x_2$  space will lead to even smaller determinants for the resulting FIM, this implies that for the given two-factor experiment, and the postulated model shown above, the  $2^2$  factorial design maximizes the FIM and hence is the “D-Optimal” design for this problem.)

## APPLICATION PROBLEMS

**19.18** The following table of data from Nelson (1989)<sup>14</sup> shows the “cold cranking power” of five different battery types quantified as the number of seconds that a particular battery generated its rated amperage without falling below 7.2 volts, at 0° F. The experiment was replicated four times for each battery type.

| Battery Type →  | 1  | 2  | 3  | 4  | 5  |
|-----------------|----|----|----|----|----|
| Experiment No ↓ | 1  | 41 | 42 | 27 | 48 |
| 1               | 41 | 42 | 27 | 48 | 28 |
| 2               | 43 | 43 | 26 | 45 | 32 |
| 3               | 42 | 46 | 28 | 51 | 37 |
| 4               | 46 | 38 | 27 | 46 | 25 |

When presented in Problem 12.22, the objective then was to identify any suggestion of “descriptive” (as opposed to “inductive”) evidence in the data set to support the postulate that some battery types are better than others.

- (i) Now, use a one-way classification ANOVA to determine inductively if there is a difference in the “cold cranking power” of these battery types. What assumptions are required for this to be a valid test? Are these assumptions reasonable?
- (ii) From a box plot of the data set, which battery types appear to be different from the others?

**19.19** Refer to Problem 19.18 and the data table. Use a computer program to carry out a nonparametric Kruskal-Wallis test. Interpret your result. What conclusion does this result lead to regarding the equality of the “cold cranking power” of these battery types? Is this conclusion different from the one reached in Problem 19.18

<sup>14</sup>Nelson, L.S., (1989). “Comparison of Poisson means,” *J. of Qual. Tech.*, 19, 173–179.

using the ANOVA method?

**19.20** Moore *et al.* (1972)<sup>15</sup>, present the following data on the weight gain (in pounds) for pigs sampled from five different litters.

| Litter Number |    |    |    |    |
|---------------|----|----|----|----|
| L1            | L2 | L3 | L4 | L5 |
| 23            | 29 | 38 | 30 | 31 |
| 27            | 25 | 31 | 27 | 33 |
| 26            | 33 | 28 | 28 | 31 |
| 19            | 36 | 35 | 22 | 28 |
| 30            | 32 | 33 | 33 | 30 |
|               | 28 | 36 | 34 | 24 |
|               | 30 | 34 | 29 |    |
|               | 31 | 32 | 30 |    |

- (i) At the  $\alpha = 0.05$  significance level, test the hypothesis that the litter from which a pig is drawn has no effect on the weight gain. What is the conclusion of this test?
- (ii) Had the test been conducted at the  $\alpha = 0.01$  significance level, what conclusions would you have reached? Examine a box plot of the data and comment on what it suggests about the hypothesis.

**19.21** The table below shows the time in months between occurrences of safety violations for three operators, “A,” “B,” and “C,” working in a toll manufacturing facility.

|   |      |      |      |      |      |      |      |      |      |      |
|---|------|------|------|------|------|------|------|------|------|------|
| A | 1.31 | 0.15 | 3.02 | 3.17 | 4.84 | 0.71 | 0.70 | 1.41 | 2.68 | 0.68 |
| B | 1.94 | 3.21 | 2.91 | 1.66 | 1.51 | 0.30 | 0.05 | 1.62 | 6.75 | 1.29 |
| C | 0.79 | 1.22 | 0.65 | 3.90 | 0.18 | 0.57 | 7.26 | 0.43 | 0.96 | 3.76 |

The data is clearly *not* normally distributed since the phenomenon in question is typical of an exponential random variable. Carry out an appropriate test to validate the hypothesis that there *is* a difference between the safety performance of these operators. Justify your test choice and interpret your results adequately.

**19.22** The following table, adapted from Gilbert (1973)<sup>16</sup>, shows the sprinting speeds (in feet/second) for three types of fast animals, classified by sex (male or female). Analyze the data appropriately. From your analysis, what can you conclude about the “effect” of animal type and sex on speed?

<sup>15</sup>P.G. Moore, A.C. Shirley, and D.E. Edwards, (1972). *Standard Statistical Calculations*, Pitman, Bath.

<sup>16</sup>Gilbert, (1973). *Biometrical Interpretation*, Clarendon Press, Oxford.

| Fast Animal Sprinting Speeds |         |           |          |
|------------------------------|---------|-----------|----------|
| Animal Type →<br>Sex ↓       | Cheetah | Greyhound | Kangaroo |
| M                            | 56      | 37        | 44       |
|                              | 52      | 33        | 48       |
|                              | 55      | 40        | 47       |
| F                            | 53      | 38        | 39       |
|                              | 50      | 42        | 41       |
|                              | 51      | 41        | 36       |

**19.23** A study of the effects of 3 components  $A$ ,  $B$ , and  $C$  on the etched grid line width in a photolithography process was carried out using a  $2^3$  factorial design with the following levels of each factor:

| Factor | Low | High |
|--------|-----|------|
| $A$    | 0%  | 6%   |
| $B$    | 8%  | 16%  |
| $C$    | 0%  | 3%   |

Generate a standard (un-replicated) design and analyze the following results for the response  $y$  = grid line width (in coded units) obtained in the series of 8 experiments run in randomized order, but which, for the purpose of this problem, have been rearranged in the standard order as follows: 6.6, 7.0, 9.7, 9.4, 7.2, 7.6, 10.0, and 9.8. Which effects are significant?

**19.24** Box and Bisgaard, (1987)<sup>17</sup>, present an experimental study of a process for manufacturing carbon-steel springs. The study was designed to identify process operating conditions that will minimize or possibly eliminate the cracks that have plagued the manufactured springs. From physics, material science, and process knowledge, the following factors and levels were chosen for the investigation, along with a  $2^3$  factorial design with no replication, for a grand total of 8 experiments. The response of interest,  $y$ , is the proportion (in %) of the manufactured batch of springs that *do not crack*.

| Factor                                    | Low  | High |
|---|------|------|
| $x_1$ : Steel Temperature ( $^{\circ}$ F) | 1450 | 1600 |
| $x_2$ : Carbon Content (%)                | 0.5  | 0.7  |
| $x_3$ : Quench Oil Temp. ( $^{\circ}$ F)  | 70   | 120  |

The result of the study is shown in the data table below, presented in standard order (the actual experiments were run in random order).

- (i) Analyze the experimental results to determine which factors and interactions are significant. State clearly how you were able to determine significance, complete with standard errors of the estimated significant main effects and interactions.
- (ii) Assess the model validity.
- (iii) Interpret the results of this analysis in terms of what should be done if the objective is to increase the percentage of manufactured springs that do not crack.

<sup>17</sup>Box, G.E.P. and S. Bisgaard, (1987). "The scientific context of quality improvement," *Quality Progress*, 22 (6) 54–61.

| Run | $x_1$ | $x_2$ | $x_3$ | $y$ |
|-----|-------|-------|-------|-----|
| 1   | -1    | -1    | -1    | 67  |
| 2   | 1     | -1    | -1    | 79  |
| 3   | -1    | 1     | -1    | 61  |
| 4   | 1     | 1     | -1    | 75  |
| 5   | -1    | -1    | 1     | 59  |
| 6   | 1     | -1    | 1     | 90  |
| 7   | -1    | 1     | 1     | 52  |
| 8   | 1     | 1     | 1     | 87  |

**19.25** Palamakula, *et al.* (2004)<sup>18</sup>, carried out a study to determine a model to use in optimizing capsule dosage for a highly lipophilic compound, Coenzyme Q10. The study involved 3 factors, Limonene, Cremophor, and Capmul, in a Box-Behnken design; the primary response was the cumulative percentage of drug released after 5 minutes (even though the paper reported 4 other responses which served merely as constraints). The low and high settings for each variable are shown below (the middle setting is exactly halfway in between).

| Factor             | Low | High |
|--------------------|-----|------|
| Limonene (mg)      | 18  | 81   |
| Cremophor EL (mg)  | 7.2 | 57.6 |
| Capnmul GMO50 (mg) | 1.8 | 12.6 |

The results of interest (presented in standard order) are shown in the table below. Generate a design, analyze the results and decide on a model that contains only statistically significant parameters. Justify your decision. After you have completed your analysis, compare your findings with those reported in the paper.

| Run<br>Std<br>Order | Response<br>$y \pm SD$ |
|---------------------|------------------------|
| 1                   | $44.4 \pm 29.8$        |
| 2                   | $6 \pm 9.82$           |
| 3                   | $3.75 \pm 6.5$         |
| 4                   | $1.82 \pm 1.07$        |
| 5                   | $18.2 \pm 8.73$        |
| 6                   | $57.8 \pm 9.69$        |
| 7                   | $68.4 \pm 1.65$        |
| 8                   | $3.95 \pm 3.63$        |
| 9                   | $58.4 \pm 2.56$        |
| 10                  | $24.8 \pm 5.80$        |
| 11                  | $1.60 \pm 1.49$        |
| 12                  | $12.1 \pm 0.84$        |
| 13                  | $81.2 \pm 9.90$        |
| 14                  | $72.1 \pm 7.32$        |
| 15                  | $82.06 \pm 10.2$       |

**19.26** The following data set is from an experimental study reported in Garge,

<sup>18</sup>Palamakula, A., M.T.H. Nutan and M. A. Khan (2004). "Response Surface Methodology for optimization and characterization of limonene-based Coenzyme Q-10 self-nanoemulsified capsule dosage form." *AAPS PharmSciTech*, 5 (4), Article 66. (Available at <http://www.aapspharmscitech.org/articles/pt0504/pt050466/pt050466.pdf>.)

(2007)<sup>19</sup>, designed to understand the complex mechanisms involved in the reactive extrusion process. The primary variables of interest (which define the extruder operating conditions) are: melting zone barrel temperature, mixing zone temperature, feed rate, screw speed, base feed composition and pulse composition. The response variables are the Melting Energy (J) and Reaction Energy (J). To determine which factors have statistically significant effects on the responses, experiments were conducted using a resolution IV,  $2^{6-2}$  fractional factorial design, with the high and low settings for the factors chosen to ensure stable extruder operation under each operating condition. The design is shown below.

| Run | Melting Zone Temp (°C) | Mixing Zone Temp (°C) | Inlet composition (% E) | Screw Speed (RPM) | Feed Rate (lb/h) | Pulse Composition (% E) |
|-----|------------------------|-----------------------|-------------------------|-------------------|------------------|-------------------------|
| 1   | 135                    | 135                   | 5                       | 150               | 25               | 100                     |
| 2   | 135                    | 135                   | 5                       | 250               | 25               | 10                      |
| 3   | 135                    | 135                   | 0                       | 250               | 15               | 100                     |
| 4   | 150                    | 150                   | 5                       | 250               | 25               | 100                     |
| 5   | 135                    | 150                   | 0                       | 150               | 25               | 10                      |
| 6   | 150                    | 150                   | 0                       | 250               | 15               | 10                      |
| 7   | 135                    | 150                   | 0                       | 250               | 25               | 100                     |
| 8   | 135                    | 135                   | 0                       | 150               | 15               | 10                      |
| 9   | 150                    | 150                   | 5                       | 150               | 25               | 10                      |
| 10  | 150                    | 135                   | 5                       | 150               | 15               | 100                     |
| 11  | 150                    | 135                   | 0                       | 250               | 25               | 100                     |
| 12  | 135                    | 150                   | 5                       | 150               | 15               | 10                      |
| 13  | 150                    | 135                   | 5                       | 250               | 15               | 100                     |
| 14  | 135                    | 150                   | 5                       | 250               | 15               | 100                     |
| 15  | 150                    | 135                   | 0                       | 150               | 25               | 10                      |
| 16  | 150                    | 150                   | 0                       | 150               | 15               | 100                     |

The results of the experiments (run in random order, but presented in standard order) are shown in the table below. For each response, analyze the data and determine which factors are significant. Comment on the model fit.

<sup>19</sup>S. Garge, (2007). "Development of an inference-based control scheme for reactive extrusion processes," PhD Dissertation, University of Delaware.

| Run | Melting Energy (J) | Reaction Energy (J) |
|-----|--------------------|---------------------|
| 1   | 1700               | 1000                |
| 2   | 1550               | 300                 |
| 3   | 1300               | 3700                |
| 4   | 800                | 800                 |
| 5   | 800                | 100                 |
| 6   | 650                | 0                   |
| 7   | 650                | 800                 |
| 8   | 650                | 0                   |
| 9   | 1100               | 0                   |
| 10  | 1000               | 600                 |
| 11  | 1650               | 650                 |
| 12  | 650                | 100                 |
| 13  | 1100               | 2100                |
| 14  | 650                | 1400                |
| 15  | 1300               | 0                   |
| 16  | 950                | 2150                |

# Chapter 20

---

## *Application Case Studies III: Statistics*

|        |   |     |
|--------|---|-----|
| 20.1   | Introduction .....  | 856 |
| 20.2   | Prussian Army Death-by-Horse kicks .....                      | 857 |
| 20.2.1 | Background and Data .....                                     | 857 |
| 20.2.2 | Parameter Estimation and Model Validation .....               | 858 |
| 20.2.3 | Recursive Bayesian Estimation .....                           | 860 |
|        | Motivation, Background and Data .....                         | 860 |
|        | Theory: The Bayesian MAP Estimate .....                       | 862 |
|        | Application: Recursive Bayesian Estimation Formula .....      | 864 |
|        | Application: Recursive Bayesian Estimation Results .....      | 866 |
|        | Final Remarks .....   | 867 |
| 20.3   | WW II Aerial Bombardment of London .....                      | 868 |
| 20.4   | US Population Dynamics: 1790-2000 .....                       | 870 |
| 20.4.1 | Background and Data .....                                     | 870 |
| 20.4.2 | “Truncated Data” Modeling and Evaluation .....                | 872 |
| 20.4.3 | Full Data Set Modeling and Evaluation .....                   | 873 |
|        | Future Prediction .....                                       | 874 |
|        | Hypothesis Testing Concerning Average Population Growth Rate  | 876 |
| 20.5   | Process Optimization .....                                    | 878 |
| 20.5.1 | Problem Definition and Background .....                       | 879 |
| 20.5.2 | Experimental Strategy and Results .....                       | 879 |
|        | Planning .....  | 879 |
|        | Design and Implementation .....                               | 879 |
| 20.5.3 | Analysis .....  | 880 |
|        | Optimization .....  | 883 |
|        | Confirmation .....  | 889 |
| 20.6   | Summary and Conclusions .....                                 | 889 |
|        | PROJECT ASSIGNMENTS .....                                     | 890 |
|        | 1. Effect of Bayesian Prior Distributions on Estimation ..... | 890 |
|        | 2. “First Principles” Population Dynamics Modeling .....      | 890 |
|        | 3. Experimental Design and Analysis .....                     | 891 |
|        | 4. Process Development and Optimization .....                 | 891 |

*A good edge is good for nothing  
if it has nothing to cut.*

Thomas Fuller (1608–1661)

As most seasoned practitioners know, there is a certain “art” to the analysis of real data. It is not just that real life never quite seems to fit nicely into the neat little ideal boxes of our theoretical constructs; it is also all the little (and not so little) wrinkles unique to each problem that make data analysis what it is. Simplifying assumptions must be made, and strategies must be formulated

for how best to approach the problem; and what works in one case may not necessarily work in another. But even art has its foundational principles, and each art form its own peculiar set of tools. Whether it is capturing the detail in a charcoal-on-paper portrait of a familiar face, the deep perspectives in an oil-on-canvas painting of a vast landscape, or the rugged three-dimensional contours of a sculpture, the problem at hand is what ultimately recommends the tools. Such is the case with the three categories of problems selected for discussion in this final chapter in the trilogy of case studies. They have been selected to demonstrate the broad range of applications of the theory and principles of the past few chapters.

The first problem is actually a pair of distinct but similarly structured problems, staples of introductory probability and statistics courses that are frequently used to demonstrate the powers of the Poisson distribution in capturing the elusive character of rare events. The first, the famous von Bortkiewicz data set on death-by-horse kicks in the 19<sup>th</sup> century Prussian army, involves an “after-the-fact” analysis of unusual death; the second, an analysis of the aerial bombardment of London in the middle of World War II, is marked by a “cannot-waste-a-minute” urgency of the need to protect the living. One is pure analysis (to which we have added a bit of a wrinkle); the other is a brilliant use of analysis and hypothesis testing for practical life-and-death decision making.

The second problem involves the complete 21 decades of US census data from 1790 to 2000. By the standards of sheer volume, this is a comparatively modest data set (especially when compared to the data sets in the first problem that are at least an order of magnitude larger). For a data set that could be analyzed in an almost limitless number of ways, it is interesting, as we show here, how a simple regression analysis, an examination of the residuals, and other such investigations can provide glimpses (smudges?) of the fingerprints left by history on a humble data set consisting of only 22 entries.

The third problem is more prosaic, coming from the no-nonsense world of industrial manufacturing. It involves process optimization, using strategic experimental designs and data analysis to meet difficult business objectives. But what this problem lacks in the macabre drama of the first, or in the rich history of the second, it tries to make up for in hard-headed practicality.

Which of these three sets of problems is the charcoal-on-paper portrait, the oil-on-canvas landscape, or the rugged three-dimensional sculpture is left to the reader’s imagination.

---

## 20.1 Introduction

Having completed our intended course of discussion on descriptive, inferential, and experimental design aspects of statistics, the primary objective of

this chapter is to present a few real-life problems to demonstrate how the concepts and ideas discussed in the preceding chapters have been (and continue to be) used to find appropriate solutions to important problems. The following is a brief catalog of the problems selected for discussion in this chapter, along with what aspects of statistics they illustrate:

1. The *Prussian army death-by-horse kicks* problem involves the analysis of the truly rare events implied in the title. It illustrates probability modeling, the characterization of a population using the concepts of sampling and estimation, and illustrates probability model validation. Because the data set is a 20-year record of events happening year-by-year, we add a wrinkle to this famous problem by investigating what could have happened had recursive Bayesian estimation been used to analyze the data, not all at once at the end of the 20-year period, but year-to-year. The question “what is such a model useful for?” is answered by a complementary problem, similar in structure but different in detail.
2. The *Aerial Bombardment of London in World War II* problem provides an answer to the question of “practicality” raised by the Poisson modeling and analysis of the Prussian army data. This latter problem picks up where former one left off, by demonstrating how a Poisson model and an appropriately framed hypothesis test provided the basis for the strategic deployment of scarce anti-aircraft resources. In the historical context of the time, solving this latter problem was anything but an act of mere intellectual curiosity.
3. The *US Population dynamics* problem illustrates the power of simple regression modeling, and how it can be used judiciously to make predictions. It also illustrates how data analysis can be used almost forensically to find hidden clues in data sets.
4. The *Process Optimization* problem illustrates the use of design of experiments (especially response surface designs) to find optimum conditions for achieving manufacturing objectives in an industrial process.

---

## 20.2 Prussian Army Death-by-Horse kicks

### 20.2.1 Background and Data

In 1898, Ladislaus von Bortkiewicz, a Russian economist and statistician of Polish origin, published an important paper in the history of probability modeling of random phenomena and statistical data analysis<sup>1</sup>. The original

---

<sup>1</sup>L. von Bortkiewicz, (1898). *Das Gesetz der Kleinen Zahlen*, Leipzig, Teubner.

**TABLE 20.1:** Frequency distribution of Prussian army deaths by horse kicks

| No of Deaths<br><i>x</i> | Number of occurrences<br>of <i>x</i> deaths per unit-year<br>(Total Frequency) |
|--------------------------|--|
| 0                        | 109  |
| 1                        | 65   |
| 2                        | 22   |
| 3                        | 3  |
| 4                        | 1  |
| Total                    | 200  |

paper contained data from a 20-year study, from 1875–1894, of 14 Prussian cavalry units, recording how many members of each cavalry unit died from a horse kick. Table 20.1 shows the popularized version of the data set, the frequency distribution of 200 observations (data from 10 units in 20 years). It is based on 10 of the 14 corps, after R.A. Fisher (of the *F*-distribution, and ANOVA) in 1925 removed data from 4 corps because they were organized differently from the others.

In the original publication, von Bortkiewicz used the data to demonstrate that rare events (with low probability of occurrence), when considered in large populations, tend to follow the Poisson distribution. As such, if the random variable,  $X$ , represents the total number of deaths  $0, 1, 2, 3, \dots$  recorded by the Prussian army over this period, it should be very well-modeled as a Poisson random variable with the pdf:

$$f(x) = \frac{\theta^x e^{-\theta}}{x!} \quad (20.1)$$

where the parameter  $\theta$  is the mean number of deaths recorded per unit-year.

Observe that the phenomena underlying this problem fit those stated for the Poisson random variable in Chapter 8: the variable of interest is the total number of occurrences of a rare event; the events are occurring in a fixed interval of time (and location); and they are assumed to occur at a uniform average rate. The primary purpose here is twofold: to characterize this random variable,  $X$ , by determining the defining population parameter (in this case, the Poisson mean number of deaths per unit-year), and to confirm that the model is appropriate.

The data shown above is clearly from an “observational” study. No one designed the experiment, *per se* (how could one?); the deaths were simply recorded each year for each cavalry unit as they occurred. Nevertheless, there is no evidence to suggest that the horses and their victims came in contact in any other way than randomly. It is therefore reasonable to consider the data as a random sample from this population of 19<sup>th</sup> century cavalry units.

**TABLE 20.2:** Actual vs Predicted Frequency distribution of Prussian army deaths

| No of Deaths<br>$x$ | Number of occurrences<br>of $x$ deaths per unit-year<br>(Total Frequency) | Predicted<br>Total Frequency |
|---------------------|---|------------------------------|
| 0                   | 109   | 108.7                        |
| 1                   | 65  | 66.3                         |
| 2                   | 22  | 20.2                         |
| 3                   | 3   | 4.1                          |
| 4                   | 1   | 0.6                          |
| 5                   | 0   | 0.1                          |
| $\geq 6$            | 0   | 0.0                          |
| Total               | 200   | 200                          |

### 20.2.2 Parameter Estimation and Model Validation

If the data set is considered a random sample, then the maximum likelihood estimate of the parameter,  $\theta$ , is the sample average, which is obtained from the data presented in Table 20.1 as:

$$\begin{aligned}\bar{x} &= \frac{(0 \times 109) + (1 \times 65) + (2 \times 22) + (3 \times 3) + (4 \times 1)}{200} \\ &= \frac{122}{200} = 0.61\end{aligned}\tag{20.2}$$

Thus, the point estimate is

$$\hat{\theta} = 0.61\tag{20.3}$$

and with sample variance  $s^2 = 0.611$ , the standard error of the estimate is

$$SE(\theta) = 0.041\tag{20.4}$$

The 95% confidence interval estimate is therefore

$$\hat{\theta} = 0.61 \pm 1.96 \times 0.041 = 0.61 \pm 0.081\tag{20.5}$$

From the point estimate, the *predicted* relative frequency distribution,  $\hat{f}(x)$ , is therefore obtained according to:

$$\hat{f}(x) = f(x|\theta = 0.61) = \frac{0.61^x e^{-0.61}}{x!}\tag{20.6}$$

The corresponding predicted total frequencies (for  $n = 200$ ), may now be computed as  $n\hat{f}(x)$ ; the result is shown alongside the original data in Table 20.2, from where the agreement between data and model is seen to be quite good.

To quantify how well the Poisson model fits the data, we may now carry out the Chi-squared goodness-of-fit test using MINITAB (recall the discussion

in Chapter 17). The result is shown below.

#### Goodness-of-Fit Test for Poisson Distribution

Data column: X

Frequency column: Actual

Poisson mean for X = 0.61

| X   | Poisson  |             | Contribution |                |
|-----|----------|-------------|--------------|----------------|
|     | Observed | Probability | Expected     | to Chi-Sq      |
| 0   | 109      | 0.543351    | 108.670      | 0.001001       |
| 1   | 65       | 0.331444    | 66.289       | 0.025057       |
| 2   | 22       | 0.101090    | 20.218       | 0.157048       |
| >=3 | 4        | 0.024115    | 4.823        | 0.140417       |
|     |          | N           | DF           | Chi-Sq P-Value |
|     |          | 200         | 2            | 0.323524 0.851 |

1 cell(s) (25.00%) with expected value(s) less than 5.

With a computed chi-squared statistic,  $C^2 = 0.323$ , being so small, and an associated  $p$ -value of 0.851, i.e.,

$$P(\chi^2(2) > 0.323) = 0.851 \quad (20.7)$$

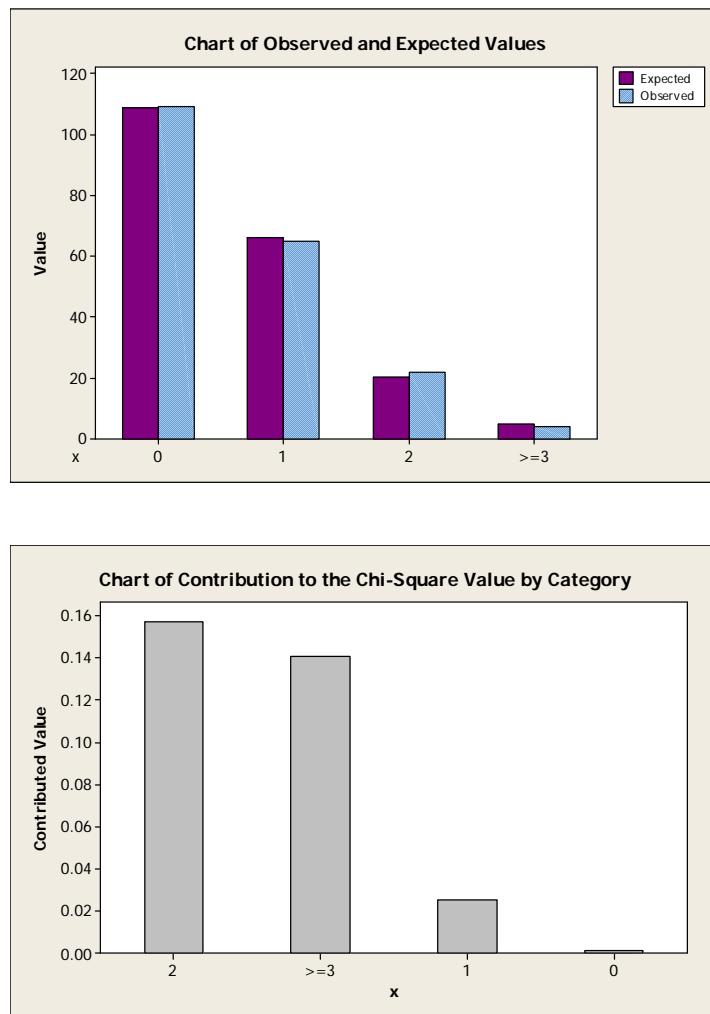
we have no evidence to support rejecting the null hypothesis (at the significance level of 0.05) and therefore conclude that the model provides an adequate fit to the data. Note that the last two frequency groups corresponding to  $X = 3$  and  $X = 4$  had to be combined for the chi-squared test; and even then, the expected frequency,  $n\hat{f} = 4.823$  fell just short of the required 5. MINITAB identifies this and prints out a warning. However, this is not enough to invalidate the test.

A graphical representation of the observed versus predicted frequency, along with a bar graph of the individual contributions to the chi-squared statistic from each frequency group is shown in Fig 20.1.

#### 20.2.3 Recursive Bayesian Estimation

##### Motivation, Background and Data

As good a fit as the Poisson model provided to the Prussian army data, it should not be lost on the reader that it took 20 years to accumulate the data. That it took so long to acquire this now famously useful data set is understandable: by definition, stable analysis of such rare-events phenomena requires the accumulation of sufficient data to allow enough occurrences of these rare events. This brings up an interesting idea: is it possible to obtain parameter estimates sequentially from this data set as it is being built up, from year-to-year, so that analysis need not wait until *all* the data is available in its entirety? The obvious risk is that such year-to-year estimates will be unreliable



**FIGURE 20.1:** Chi-Squared test results for Prussian army death by horse kicks data and a postulated Poisson model. Top panel: Bar chart of "Expected" and "Observed" frequencies; Bottom Panel: Bar chart of contributions to the Chi-squared statistic.

**TABLE 20.3:** Year-by-Year, Unit-by-Unit breakdown of Prussian army deaths data

| Unit →<br>Year ↓ | 1 | 2 | 3 | 4 | 5 | 6 | 6 | 8 | 9 | 10 | Total |
|------------------|---|---|---|---|---|---|---|---|---|----|-------|
| 1                | 1 | 0 | 2 | 0 | 0 | 0 | 1 | 0 | 0 | 0  | 4     |
| 2                | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0  | 3     |
| 3                | 0 | 1 | 0 | 0 | 2 | 0 | 1 | 2 | 0 | 0  | 6     |
| 4                | 0 | 1 | 3 | 1 | 0 | 0 | 1 | 0 | 0 | 1  | 7     |
| 5                | 0 | 0 | 2 | 0 | 1 | 4 | 0 | 2 | 1 | 0  | 10    |
| 6                | 0 | 0 | 2 | 0 | 1 | 0 | 2 | 1 | 0 | 1  | 7     |
| 7                | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1  | 7     |
| 8                | 0 | 1 | 0 | 2 | 0 | 0 | 0 | 1 | 0 | 0  | 5     |
| 9                | 1 | 0 | 1 | 0 | 1 | 0 | 3 | 0 | 1 | 1  | 8     |
| 10               | 0 | 0 | 2 | 0 | 1 | 0 | 0 | 1 | 0 | 0  | 4     |
| 11               | 1 | 1 | 0 | 3 | 0 | 2 | 1 | 0 | 0 | 0  | 8     |
| 12               | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 2 | 0  | 5     |
| 13               | 1 | 0 | 0 | 0 | 2 | 0 | 1 | 1 | 2 | 2  | 9     |
| 14               | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0  | 2     |
| 15               | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0  | 5     |
| 16               | 1 | 0 | 2 | 1 | 0 | 1 | 1 | 1 | 0 | 1  | 8     |
| 17               | 0 | 1 | 0 | 2 | 1 | 0 | 2 | 0 | 1 | 1  | 8     |
| 18               | 0 | 1 | 2 | 2 | 0 | 1 | 0 | 0 | 0 | 0  | 6     |
| 19               | 2 | 0 | 1 | 0 | 0 | 2 | 1 | 0 | 0 | 1  | 7     |
| 20               | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0  | 3     |

because they will be based on sparse and information-deficient data, at least until sufficient, reliably stable information has accumulated in the growing data set. Still, is it even possible to obtain such sequential estimates? And if so, what trajectory will these “interim” estimates take in their approach to the final full-data set estimate?

The answer to the first question is yes; and Bayesian estimation is one technique for implementing such a sequential estimation strategy. And to provide a specific answer to the second question, we now present a Bayesian procedure for sequentially estimating the Poisson parameter, simulating how the analysis *might have progressed* “in real-time,” year-by-year, if one imagined that the data became available every year as shown in the Table 20.3 for the 10 units.

### Theory: The Bayesian MAP Estimate

Bayesian estimation, as we may recall, requires that we first specify a prior distribution to be combined with the sampling distribution to obtain the posterior distribution, from which the parameter estimates are then obtained. In Chapter 14, we recommended the use of conjugate priors wherever possible. For this particular problem involving a Poisson random variable, the Gamma distribution is the conjugate prior for estimating the Poisson parameter. Thus, if we consider as a prior distribution for this unknown parameter,  $\theta$ , the gamma

$\gamma(a, b)$  distribution, i.e.,

$$f(\theta) = \frac{1}{b^a \Gamma(a)} \theta^{a-1} e^{-\theta/b} \quad (20.8)$$

we can then obtain the posterior distribution,  $f(\theta|x_1, x_2, \dots, x_n)$ , after combining this prior with the sampling distribution for a random sample, drawn from the Poisson distribution. From this posterior distribution, we will use the maximum *a-posteriori* (MAP) estimate as our choice for the parameter estimate.

Now, the sampling distribution for a random sample,  $X_1, X_2, \dots, X_n$ , from the Poisson distribution is given by:

$$f(x_1, x_2, \dots, x_n | \theta) = \frac{\theta^{\sum_{i=1}^n x_i} e^{-n\theta}}{x_1! x_2! \dots x_n!} \quad (20.9)$$

By Bayes' theorem, using the given prior, the posterior distribution is:

$$f(\theta|x_1, x_2, \dots, x_n) = C \frac{1}{b^a \Gamma(a) x_1! x_2! \dots x_n!} \theta^{(a-1+\sum_{i=1}^n x_i)} e^{-\theta(n+1/b)} \quad (20.10)$$

where  $C$  is the usual normalizing constant. The MAP estimate is obtained from here by taking derivatives with respect to  $\theta$ , equating to zero and solving; the same results are obtained by maximizing  $\ln f$ ; i.e.,

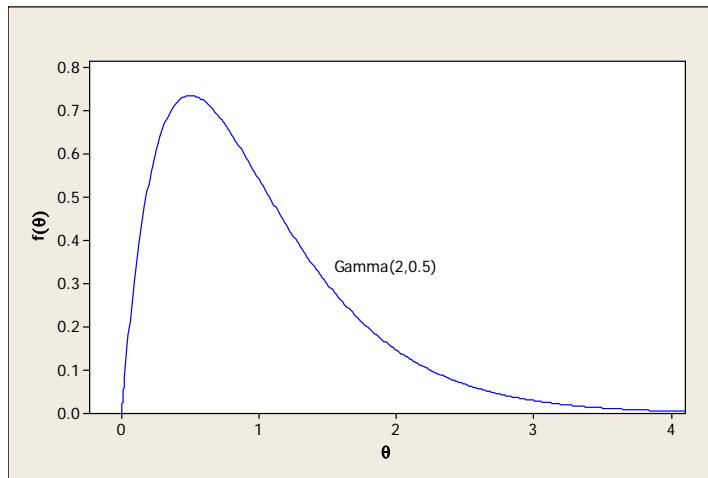
$$\frac{\partial \ln f}{\partial \theta} = \frac{1}{\theta} \left( a - 1 + \sum_{i=1}^n x_i \right) - \left( n + \frac{1}{b} \right) \quad (20.11)$$

which, upon equating to zero and solving, gives the result:

$$\hat{\theta}_{MAP} = \frac{\left( \sum_{i=1}^n x_i \right) + (a-1)}{\left( n + \frac{1}{b} \right)} \quad (20.12)$$

Of course, depending on the value specified for  $a$  and  $b$  in the prior distribution, Eq (20.12) will return different estimates for the unknown parameter. In particular, observe that for  $a = 1$  and  $b = \infty$ , the result is precisely the same as the sample average. This is one of the main criticisms of Bayesian estimation: that the subjectivity inherent in the choice of the prior distribution introduces bias into the estimate.

However, it can be shown that this bias is traded off for a smaller parameter estimate variance. Furthermore, when used recursively, whereby the posterior distribution at the current time is used as the prior distribution in the next round, the bias introduced at the beginning by the original prior distribution progressively “washes out” with each iteration. Thus, in return for the initial bias, this recursive strategy allows one to carry out estimation sequentially without having to wait for the full data set to be completely accumulated, obtaining more reliable estimates along the way.



**FIGURE 20.2:** Initial prior distribution, a Gamma (2,0.5), used to obtain a Bayesian estimate for the Poisson mean number of deaths per unit-year parameter.

It is possible to show that if the data sequence is “stable,” the Bayesian estimates,  $\hat{\theta}^{(j)}$ , obtained at the  $j^{th}$  iteration, will converge to the true value of  $\theta$  in the limit as  $j \rightarrow \infty$ . We now proceed to show, using the data in Table 20.3, just how quickly the Bayesian estimate converges to the true estimate of the mean number of deaths per unit-year.

#### Application: Recursive Bayesian Estimation Formula

Let us consider that at the beginning of the data gathering, knowing nothing more than what may have transpired in the past, it seems reasonable to suppose that the mean number of deaths per unit-year will be a number that lies somewhere between 0 and 4 (few units report deaths exceeding 4 in a year and, since these manner of deaths are rare events, many units will report no deaths). As a result, we start by selecting a prior distribution, Gamma (2, 0.5), whose pdf is shown in Fig 20.2. From the plot, we observe that this prior indeed expresses the belief that the unknown mean number of deaths per unit-year lies somewhere between 0 and 4. The distribution is fairly broad, indicating our fairly substantial level of uncertainty about this unknown parameter. The probability assigned to the higher numbers may be small, in keeping with common sense assessment of the phenomenon, but these probabilities are not zero.

Next, given this prior and the 10 data points for the first year in the first row of the data table,

$$x_1 = 1, x_2 = 0, \dots, x_{10} = 0$$

we can now use Eq (20.12) to obtain  $\hat{\theta}^{(1)}$ , the MAP estimate at the end of the

first year. With  $a = 2, b = 0.5$ , and  $\sum_{i=1}^{10} x_i = 4$ , we obtain

$$\hat{\theta}^{(1)} = \frac{5}{12} = 0.4167 \quad (20.13)$$

as the MAP estimate at the end of the first year.

In general, if  $\hat{\theta}^{(k)}$  is the MAP estimate obtained after the  $k^{th}$  year, (i.e., after a total of  $k$ -years worth of data), it is possible to establish that:

$$\hat{\theta}^{(k+1)} = \hat{\theta}^{(k)} + \left( \frac{10}{10k + 10 + b'} \right) [\bar{x}^{(k+1)} - \hat{\theta}^{(k)}] \quad (20.14)$$

where  $b' = 1/b$ , and  $\bar{x}^{(k+1)}$  is the arithmetic average of the 10 numbers constituting the year  $(k+1)$  data, defined by:

$$\bar{x}^{(k+1)} = \frac{1}{10} \sum_{i=10k+1}^{10k+10} x_i \quad (20.15)$$

This result is established as follows: From the general expression in Eq (20.12), we obtain that  $\hat{\theta}^{(k+1)}$  is given by:

$$\hat{\theta}^{(k+1)} = \frac{\left( \sum_{i=1}^{10k+10} x_i \right) + a - 1}{10k + 10 + b'} \quad (20.16)$$

since, in this case, the total number of data points used by the  $(k+1)^{th}$  year is  $10k + 10$ , with  $b' = 1/b$ . It is convenient to rearrange this equation as:

$$(10k + 10 + b')\hat{\theta}^{(k+1)} = \left( \sum_{i=1}^{10k+10} x_i \right) + a - 1 \quad (20.17)$$

Similarly, for the  $k^{th}$  year, we obtain

$$(10k + b')\hat{\theta}^{(k)} = \left( \sum_{i=1}^{10k} x_i \right) + a - 1 \quad (20.18)$$

Subtracting Eq (20.18) from Eq (20.17) gives:

$$(10k + 10 + b')\hat{\theta}^{(k+1)} = (10k + b')\hat{\theta}^{(k)} + \sum_{i=10k+1}^{10k+10} x_i \quad (20.19)$$

which may be rearranged to give:

$$(10k + 10 + b')\hat{\theta}^{(k+1)} = (10k + 10 + b')\hat{\theta}^{(k)} - 10\hat{\theta}^{(k)} + 10\bar{x}^{(k+1)}$$

with

$$\bar{x}^{(k+1)} = \frac{1}{10} \sum_{i=10k+1}^{10k+10} x_i \quad (20.20)$$

**TABLE 20.4:** Recursive  
(yearly) Bayesian estimates of the  
mean number of deaths per  
unit-year

| After<br>year $j$ | $\hat{\theta}^{(j)}$ | After<br>year $j$ | $\hat{\theta}^{(j)}$ |
|-------------------|----------------------|-------------------|----------------------|
| 1                 | 0.4167               | 11                | 0.6250               |
| 2                 | 0.3636               | 12                | 0.6148               |
| 3                 | 0.4375               | 13                | 0.6364               |
| 4                 | 0.5000               | 14                | 0.6056               |
| 5                 | 0.5962               | 15                | 0.5987               |
| 6                 | 0.6129               | 16                | 0.6111               |
| 7                 | 0.6250               | 17                | 0.6221               |
| 8                 | 0.6098               | 18                | 0.6209               |
| 9                 | 0.6304               | 19                | 0.6250               |
| 10                | 0.6078               | 20                | 0.6089               |

from where the required result is easily established:

$$\hat{\theta}^{(k+1)} = \hat{\theta}^{(k)} + \left( \frac{10}{10k + 10 + b'} \right) [\bar{x}^{(k+1)} - \hat{\theta}^{(k)}] \quad (20.21)$$

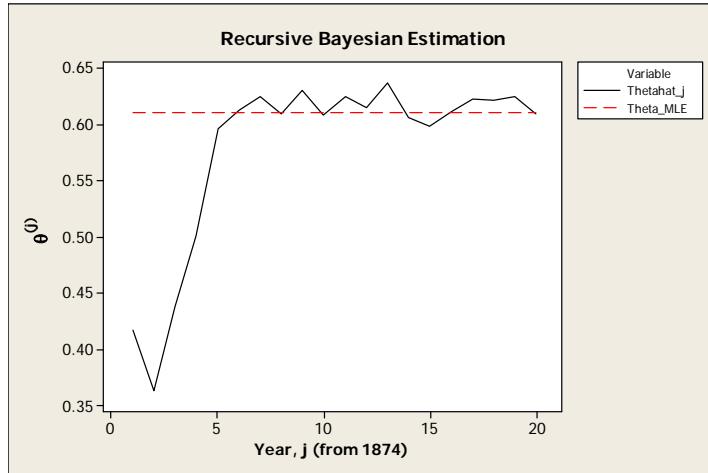
(This expression is reminiscent of the expression for the recursive least squares estimate given in Eq (16.190) in Chapter 16.)

#### Application: Recursive Bayesian Estimation Results

Beginning with  $\hat{\theta}^{(1)} = 0.4167$  obtained above, the recursive expression in Eq (20.14) may now be used along with the supplied data; the result is the sequence of MAP estimates,  $\hat{\theta}^{(j)}$ , after each  $j^{\text{th}}$  year, shown in Table 20.4 for  $j = 2, 3, \dots, 20$ .

A plot of these recursive estimates is shown in Fig 20.3 in the solid line, with the dashed line representing the single estimate (0.61) obtained earlier using the entire 20-year data all at once. Observe now that an argument could have been made for stopping the study sometime after the 5<sup>th</sup> or 6<sup>th</sup> year since, by then, the recursive estimate has essentially and observably settled down to within a small tolerance of the final value of 0.61. To be sure that enough time would have elapsed to ascertain that the value has truly settled down, the year 7 or 8 or even up to year 10 would all be good recommendations for the “stopping year.” The point is that this recursive method would have provided a stable estimate close to the final one long before the 20 years have elapsed. While the observed convergence to the maximum likelihood estimate demonstrates the “washing out” of the initial prior distribution, it will still be interesting to determine the final posterior distribution and compare it to the original prior distribution.

It can be shown (and this is left as an exercise to the reader) that the final



**FIGURE 20.3:** Recursive Bayesian estimates using yearly data sequentially, compared with the standard maximum likelihood estimate, 0.61, (dashed-line).

posterior distribution is:

$$f(\theta|x_1, x_2, \dots, x_n) = C_1 \theta^{123} e^{-202\theta} \quad (20.22)$$

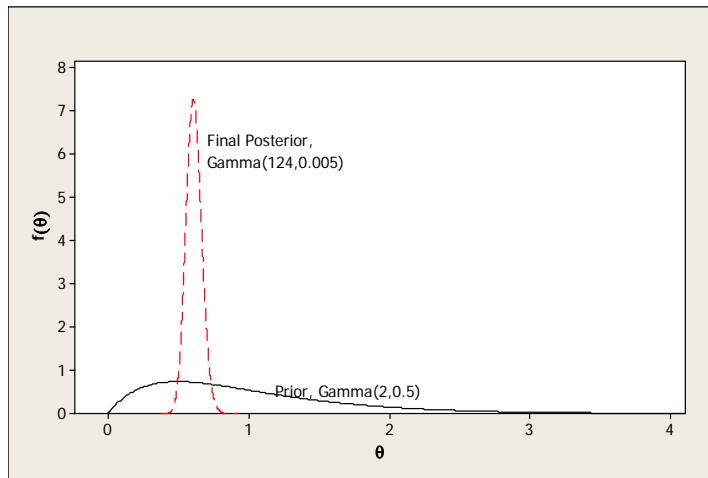
where  $C_1$  is the normalizing constant for a gamma  $\gamma(124, 1/202)$  or,

$$C_1 = \frac{202^{124}}{\Gamma(124)} \quad (20.23)$$

A plot of this distribution along with the prior gamma  $\gamma(2, 0.5)$  distribution is shown in Fig. 20.4. The key characteristic of this posterior distribution is that it is very sharply concentrated around its mean value of 0.614, and mode of 0.619, both of which are practically the same as the maximum likelihood value of 0.61 obtained earlier. This posterior distribution stands in very sharp contrast to the very broad prior distribution, demonstrating how the information obtained recursively from the accumulating data had reduced the uncertainty expressed in the initial prior distribution.

### Final Remarks

The reader will be forgiven for asking: “what is the point of this entire exercise?” After all, the soldiers whose somewhat undignified deaths (undignified, that is, for professional soldiers) have contributed to the data will not be “coming back” as a result of this exercise. Furthermore, it is not as if (at least as far as we know) the original analysis led to a change in the Prussian army policy that yielded improved protection for the soldiers against such deaths. A Poisson model fit the data well; the model parameter has been



**FIGURE 20.4:** Final posterior distribution (dashed line) along with initial prior distribution (solid line).

estimated as roughly 0.6, implying that on average, there will be 6 deaths per 10 unit-years. But what else can such a model be used for? Admittedly, the original analysis was more “just for the exercise,” to demonstrate, at the time, that the Poisson pdf is the model of choice for such phenomena. Without the historical perspective, this may appear like a pointless exercise to modern day practitioners. But with the next problem, we show how a similar Poisson model provided the crucial tool for deciding where to deploy scarce anti-aircraft artillery.

### 20.3 WW II Aerial Bombardment of London

At a critical point in World War II, after London had suffered a devastating period of sustained aerial bombardment, the British government had to make some strategic decisions about where to deploy limited anti-aircraft artillery for maximum effectiveness. How British researchers approached and solved this problem provides yet another example of how random phenomena are subject to rigorous analysis and illustrates how probability and statistics can be used effectively for solving important real-life problems.

In his 1946 paper<sup>2</sup> about this problem, R. D. Clarke noted:

---

<sup>2</sup>Clarke, R. D. (1946). “An application of the Poisson Distribution,” *J Inst. Actuaries*, 72, 48-52.

**TABLE 20.5:** Frequency distribution of bomb hits in greater London during WW II and Poisson model prediction

| No of bomb hits<br><i>x</i> | Number of regions<br>sustaining <i>x</i> bomb hits | Poisson Model<br>Prediction |
|-----------------------------|--|-----------------------------|
| 0                           | 229  | 227                         |
| 1                           | 211  | 211                         |
| 2                           | 93   | 99                          |
| 3                           | 35   | 31                          |
| 4                           | 7  | 7                           |
| 5                           | 0  | 1                           |
| 6                           | 0  | 0                           |
| 7                           | 1  | 0                           |
| Total                       | 576  | 576                         |

“During the flying-bomb attack on London, frequent assertions were made that the points of impact of the bombs tended to be grouped in clusters. It was accordingly decided to apply a statistical test to discover whether any support could be found for this allegation.”

To determine whether or not the aerial bombardment was completely haphazard with no pattern whatsoever, the greater London terrain was sub-divided into a  $24 \times 24$  grid, or 576 small square regions, each 0.25 square kilometers in size, and the number of bomb hits in each small region tallied. If the hits are random, a Poisson distribution should fit the data. Indeed, as the the data and the Poisson model prediction in Table 20.5 show, this appears to be the case.

The Poisson parameter estimate, the sample average, obtained as

$$\hat{\lambda} = 0.932 \quad (20.24)$$

is used in the model

$$\hat{f}(x) = \frac{\hat{\lambda}^x e^{-\hat{\lambda}}}{x!} \quad (20.25)$$

to obtain the model prediction shown. A Chi-squared goodness-of-fit test (not shown; left as an exercise to the reader) indeed confirms that the model is a good fit. But our task is not yet complete.

The real issue to be resolved, we must keep in mind, is whether or not certain regions were targeted for extra bombing. Identifying such regions will be crucial to the strategic deployment of anti-aircraft artillery. With this in mind, we draw attention to the entry for  $x = 7$ , indicating that one region received *seven* bomb hits. Observe that no region received 6 hits or even 5 hits; and 7 regions received four hits each (predicted perfectly by the Poisson model). This all makes one wonder: what are the chances that one region out of 576 will receive seven or more hits — purely at random?

This is where the validated model becomes useful, since we can use it to compute this particular probability as

$$P(X \geq 7 | \lambda = 0.93) = 0.000054 \quad (20.26)$$

We have retained all these decimal places to make a point. Cast in the form of a hypothesis test, this statement declares that at the 0.05 significance level, for any region to receive 7 or more hits is either the absolute rarest of rarest events, (about one in 20,000 tries) or the region in fact must have been deliberately targeted. The anti-aircraft artillery were therefore deployed in this location, and the rest is history.

## 20.4 US Population Dynamics: 1790-2000

### 20.4.1 Background and Data

The US Population—to the nearest million—as determined from census records every decade from 1790 to 2000, is shown in Table 20.6. The primary purpose of this case study is to use this data set to develop a model of how the US population has changed over the last two centuries, analyze the model for any insight, and then use it to predict the next observation in the series, the 2010 census result.

We begin by noting that this data set belongs to the category known as “time-series” data, where each observation occurs in a sequential series in time. The distinguishing characteristic of such data is the most obvious: the observations are correlated in time. If for no other reason, at least we know from natural population growth fundamentals in living organisms that the size of a population at the current time is a function of the size in the previous time period. This intrinsic correlation endows such data sets with properties that call for special “time-series analysis” tools; but these are beyond the scope of this book. Such data sets can also be analyzed using population growth models based on biology; but this requires making some fairly strong assumptions about the net population growth rate. Such assumptions may be difficult to justify, especially for a country with such complex immigration patterns as the US, and for a period spanning 210 years.

It is also possible, however, to use the tools we have discussed thus far, and this is what we plan to do in this section. For example, let us recall that a scatter plot of this very data set was shown in Chapter 12 (Fig 12.18). While the plot has the typical exponential growth characteristics of healthy populations, it also suggests that a quadratic regression model might be a good starting point. Fitting a quadratic model of the type

$$y = \theta_0 + \theta_1 x + \theta_2 x^2 + \epsilon \quad (20.27)$$

**TABLE 20.6:** US Population (to the nearest million) from 1790–2000

| Census Year | Population (millions) |
|-------------|-----------------------|
| 1790        | 4                     |
| 1800        | 5                     |
| 1810        | 7                     |
| 1820        | 10                    |
| 1830        | 13                    |
| 1840        | 17                    |
| 1850        | 23                    |
| 1860        | 31                    |
| 1870        | 40                    |
| 1880        | 50                    |
| 1890        | 63                    |
| 1900        | 76                    |
| 1910        | 92                    |
| 1920        | 106                   |
| 1930        | 123                   |
| 1940        | 132                   |
| 1950        | 151                   |
| 1960        | 179                   |
| 1970        | 203                   |
| 1980        | 227                   |
| 1990        | 249                   |
| 2000        | 281                   |

is a fairly straightforward task that we will get to in a moment. For now, because one of the objectives is to *predict* the next observation (the 2010 census result), we must be conscious of the fact that using a regression model to predict *outside* the data range is usually not a good idea.

Therefore, we propose to evaluate, first, how useful the regression model can be as a one-step and/or multiple-step ahead predictor. We plan to do this in an objective fashion as follows: we will truncate the data at 1970 and use this deliberately abbreviated data set to obtain a regression model that will then be used to predict the “missing” 1980, 1990, and 2000 population results. How well the model predictions match the actual census data will provide an objective assessment of how reasonable the regression approach can be in this regard.

#### 20.4.2 “Truncated Data” Modeling and Evaluation

Let us define a “normalized year” variable as:

$$x = \frac{\text{Census Year} - 1790}{10} + 1 \quad (20.28)$$

which essentially assigns the natural numbers 1, 2, etc to the census years, so that the first year, 1790 is 1, the second, 1800 is 2, etc. A regression analysis of the truncated data (only up to  $x = 19$ , or 1970) in MINITAB produces the following results. The regression model itself is:

$$\hat{y}_t = 6.14 - 1.86x + 0.633x^2 \quad (20.29)$$

where  $\hat{y}_t$  is the “truncated-data” model prediction, with the detailed MINITAB output shown below.

Regression Analysis: Population-t versus Xt, Xt2

The regression equation is

Population-t = 6.14 - 1.86 Xt + 0.633 Xt2

| Predictor | Coef    | SE Coef | T     | P     |
|-----------|---------|---------|-------|-------|
| Constant  | 6.137   | 2.139   | 2.87  | 0.011 |
| Xt        | -1.8630 | 0.4924  | -3.78 | 0.002 |
| Xt2       | 0.63254 | 0.02392 | 26.44 | 0.000 |

S = 2.78609 R-Sq = 99.8% R-Sq(adj) = 99.8%

These results indicate that all three parameters are significantly different from zero at the  $\alpha = 0.05$  level, and that the regression model explains a significant amount of the variability in the data — to the tune of 99.8% — without using an excessive number of parameters.

Using MINITAB’s “New Observation” prediction feature produces the following results for the model prediction of the 1980 (Obs 1), 1990 (Obs 2), and

2000 (Obs 3) census results, respectively.

| New<br>Obs | Fit   | SE Fit             | 95% CI               | 95% PI |
|------------|-------|--------------------|----------------------|--------|
| 1 221.892  | 2.139 | (217.358, 226.425) | (214.446, 229.337)X  |        |
| 2 245.963  | 2.607 | (240.437, 251.488) | (237.874, 254.051)XX |        |
| 3 271.299  | 3.131 | (264.660, 277.937) | (262.413, 280.184)XX |        |

X denotes a point that is an outlier in the predictors.

XX denotes a point that is an extreme outlier in the predictors.

Note how MINITAB flags all three predictions as “outliers” since they truly lie outside of the data range. Nevertheless, we now observe that the true values,  $y_{1980} = 227$  and  $y_{1990} = 249$ , fall nicely within the respective 95% prediction intervals, and the true value,  $y_{2000} = 281$ , falls just outside the high limit of the prediction interval. The implication is that, all things being equal, a regression model based on the full data set should be able to provide an acceptable one-step ahead prediction.

#### 20.4.3 Full Data Set Modeling and Evaluation

Repeating the entire exercise, this time using the full census data set, produces the following result. This time, the model is,

$$\hat{y} = 7.92 - 2.47x + 0.668x^2 \quad (20.30)$$

where we note that the model coefficients have not changed too drastically. The rest of the MINITAB output is shown below.

Regression Analysis: Population versus Xn, Xn2  
The regression equation is

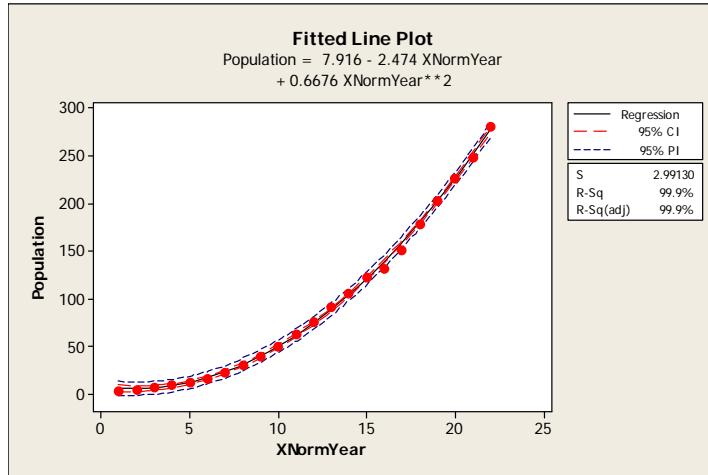
Population = 7.92 - 2.47 Xn + 0.668 Xn2

| Predictor | Coef    | SE Coef | T     | P     |
|-----------|---------|---------|-------|-------|
| Constant  | 7.916   | 2.101   | 3.77  | 0.001 |
| Xt        | -2.4735 | 0.4209  | -5.88 | 0.000 |
| Xt2       | 0.66763 | 0.01777 | 37.57 | 0.000 |

S = 2.99130 R-Sq = 99.9% R-Sq(adj) = 99.9%

Once again, the parameter estimates are seen to be significant; and the  $R^2$  and  $R^2_{adj}$  values have even improved slightly. The ANOVA table (not shown) does not show anything out of the ordinary. A plot of the data, the regression model fit, along with both the 95% confidence interval and the 95% prediction interval, is shown in Fig 20.5.

This figure seems to show that the fit is particularly good, with very little uncertainty around the model prediction, as implied by the very tight confidence and prediction intervals. However, MINITAB flagged two residuals that



**FIGURE 20.5:** Quadratic regression model fit to US Population data along with both the 95% confidence interval and the 95% prediction interval.

appear unusually large:

#### Unusual Observations

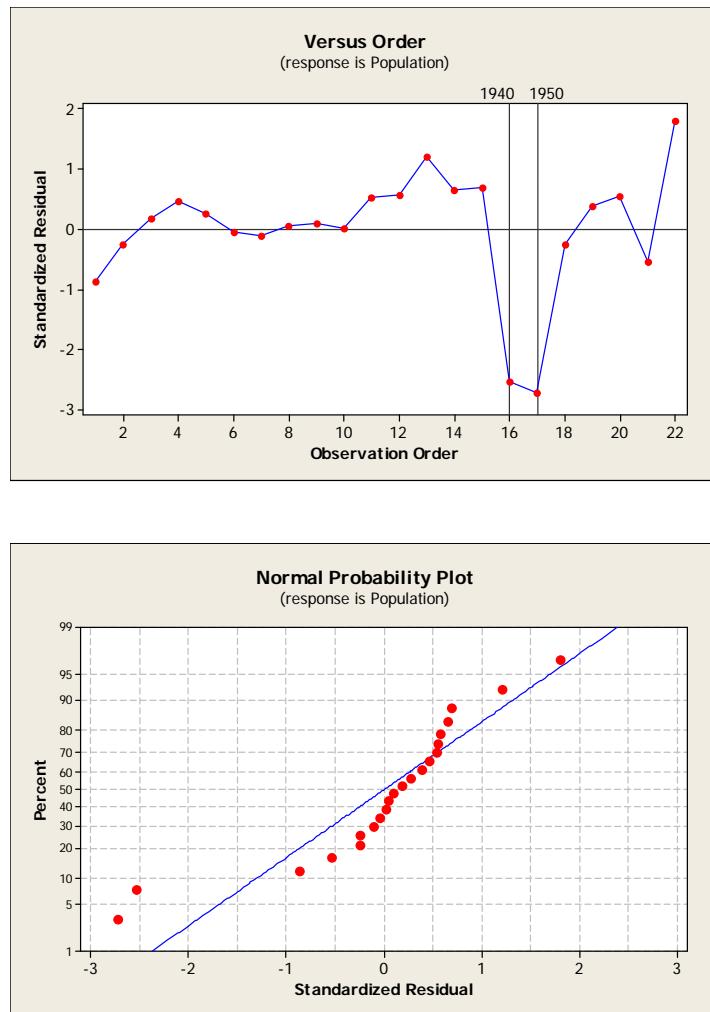
| Obs | Xn   | Population | Fit     | SE Fit | Residual | St Resid |
|-----|------|------------|---------|--------|----------|----------|
| 16  | 16.0 | 132.000    | 139.253 | 0.859  | -7.253   | -2.53R   |
| 17  | 17.0 | 151.000    | 158.811 | 0.863  | -7.811   | -2.73R   |

R denotes an observation with a large standardized residual.

The usual (standardized) residual plots shown in Fig 20.6 do in fact indicate that the residuals do not look normally distributed at all; if anything, they look serially correlated (not surprising). In particular, the two observations flagged by MINITAB are marked in the top panel of this figure. These observations belong to the census years 1940 and 1950, with the residuals indicating that the model significantly over-estimated the populations during these years (in other words, according to this model, the true population count in 1940 and in 1950 were significantly lower than expected). It is left as an exercise to the reader to suggest possible explanations for what could be responsible for a lower-than-expected population count in 1940 and 1950.

#### Future Prediction

Despite the unusual residuals, such a simple model provides a surprisingly reasonable representation of the census data. Using the model to predict the



**FIGURE 20.6:** Standardized residuals from the regression model fit to US Population data. Top panel: Residuals versus observation order; Bottom panel: Normal probability plot. Note the left-over pattern indicative of serial correlation, and the “unusual” observations identified for the 1940 and 1950 census years in the top panel; note also the general deviation of the residuals from the theoretical normal probability distribution line in the bottom panel.

2010 census result produces the following MINITAB result.

| New<br>Obs | Fit     | SE Fit | 95% CI             | 95% PI              |
|------------|---------|--------|--------------------|---------------------|
| 1          | 304.201 | 2.101  | (299.803, 308.600) | (296.550, 311.853)X |

The implication is that, to the nearest million,

$$\hat{y}_{2010} = 304; \text{ and } 297 < \hat{y}_{2010} < 312 \quad (20.31)$$

a point estimate of 304 million along with the indicated 95% prediction interval. We personally believe that this probably underestimates what the true 2010 census result will be. The potential unaccounted-for phenomena that are likely to affect this prediction include but are not limited to:

- The increasingly complex immigration patterns over the past 20 years;
- The changing mean rate of reproduction among recent immigrants and among long-term citizens;
- The influence of medical advances on life expectancy.

The reader is encouraged to think of any other potential factors that may contribute to rendering the prediction inaccurate.

#### 20.4.4 Hypothesis Testing Concerning Average Population Growth Rate

It is difficult to imagine that the average population growth rate would have remained constant from decade to decade from 1790 until modern times. To investigate this phenomenon, we generate from the census data, a table of “percent relative population growth rate” from 1800-2000 defined as follows:

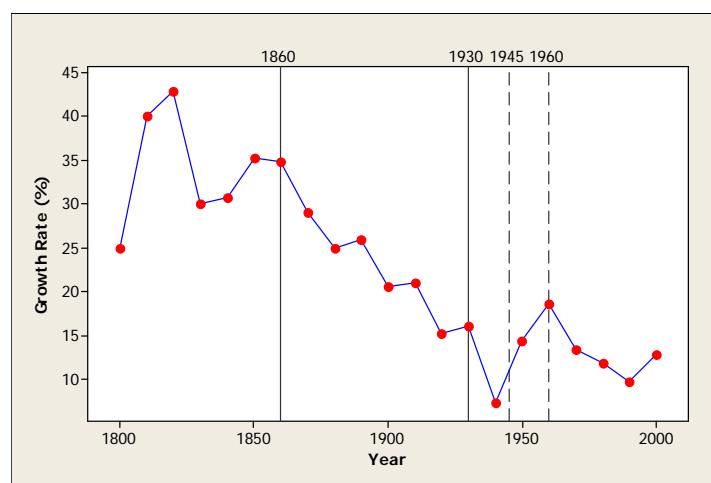
$$\Gamma(Y) = \frac{P(Y) - P(Y - 10)}{P(Y - 10)} \times 100\% \quad (20.32)$$

where  $P(Y)$  is the population recorded in year  $Y$ . The resulting 21 values may be divided into 3 even periods of 70 years each, as follows: the period from 1800-1860 is designated as “Period 1,” 1870-1930 as “Period 2,” and 1940-2000 as “Period 3.” The result is shown in Table 20.7, indicating, for example, that from 1790 to 1800, the US population experienced an average relative growth rate of 25% from 4 million to 5 million, etc.

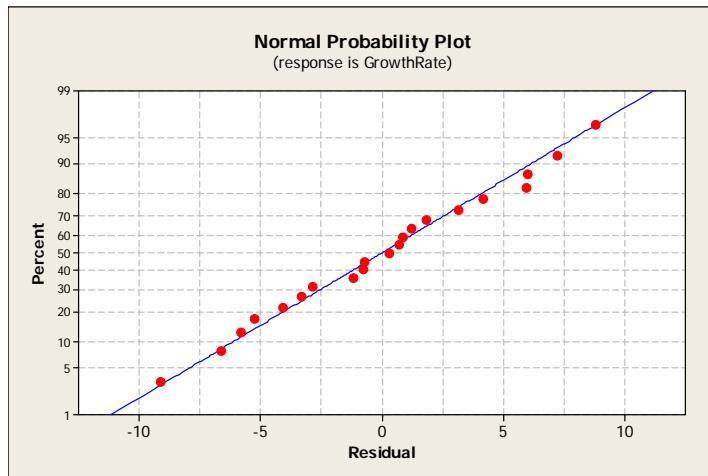
The postulate to be tested is that the average population growth rate has remained essentially constant over each of the three periods. Fig 20.7 is a plot of this relative growth rate against census year; it shows many of distinctive features, but the most obvious is that the high growth rate of the early Period 1 gave way to a decidedly sharp decline beginning in 1860, perhaps triggered by the Civil War. The decline appears to end in 1940, at the beginning of Period 3, a period marked by a steady increase in relative growth rate through 1960,

**TABLE 20.7:** Percent average relative population growth rate for each census year from 1800–2000 divided into three 70-year periods

| Census Year | Average Rel. GrowthRate (%) | Period |
|-------------|-----------------------------|--------|
| 1800        | 25.0000                     | 1      |
| 1810        | 40.0000                     | 1      |
| 1820        | 42.8571                     | 1      |
| 1830        | 30.0000                     | 1      |
| 1840        | 30.7692                     | 1      |
| 1850        | 35.2941                     | 1      |
| 1860        | 34.7826                     | 1      |
| 1870        | 29.0323                     | 2      |
| 1880        | 25.0000                     | 2      |
| 1890        | 26.0000                     | 2      |
| 1900        | 20.6349                     | 2      |
| 1910        | 21.0526                     | 2      |
| 1920        | 15.2174                     | 2      |
| 1930        | 16.0377                     | 2      |
| 1940        | 7.3171                      | 3      |
| 1950        | 14.3939                     | 3      |
| 1960        | 18.5430                     | 3      |
| 1970        | 13.4078                     | 3      |
| 1980        | 11.8227                     | 3      |
| 1990        | 9.6916                      | 3      |
| 2000        | 12.8514                     | 3      |



**FIGURE 20.7:** Percent average relative population growth rate in the US for each census year from 1800–2000 divided into three equal 70-year periods. Period 1: 1800–1860; Period 2: 1870–1930; Period 3: 1940–2000.



**FIGURE 20.8:** Normal probability plot for the residuals from the ANOVA model for Percent average relative population growth rate versus Period with Period 1: 1800-1860; Period 2: 1870-1930; Period 3: 1940-2000.

perhaps driven by the “baby boom.” The intermediate year 1945 (during WW II) and the census year, 1960, are marked on the graph for reference.

A formal one-way ANOVA test of equality of the average relative growth rates for these 3 periods shows the following not-too-surprising result:

**Results for: USPOPULATION.MTW**  
**One-way ANOVA: GrowthRate versus Period**

| Source | DF | SS     | MS    | F     | P     |
|--------|----|--------|-------|-------|-------|
| Period | 2  | 1631.9 | 816.0 | 32.00 | 0.000 |
| Error  | 18 | 458.9  | 25.5  |       |       |
| Total  | 20 | 2090.9 |       |       |       |

$$S = 5.049 \quad R-Sq = 78.05\% \quad R-Sq(\text{adj}) = 75.61\%$$

The indication is, of course, that there is a significant difference in the average relative growth rates in each period. A normal probability plot of the residuals from the one-way ANOVA model (after the “Period effects,” the average relative growth rates for each period, have been estimated) is shown in Fig 20.8. Visually, the normality assumption appears to be reasonable.

As an exercise, the reader should take a closer look at the complete “average percent population growth rate” data from 1800-2000 in Table 20.7 (and the plot in Fig 20.7), and interpret any observable features from the perspective of US History and other contemporary trends.

## 20.5 Process Optimization

### 20.5.1 Problem Definition and Background

This final problem involves improving the overall performance of a two-stage commercial coating process.<sup>3</sup> The primary material used in the coating process is produced in the first stage where manufacturing “yield” (measured in %) is the key response variable of interest. In the second stage, other additives are compounded with the primary material and the coating process completed; the primary response variable is “adhesion,” measured in grams.

To meet customer specifications and remain profitable requires yields of 91% or greater and adhesion greater than 45 grams. However, the process consistently failed to meet these objectives, and an experimental program was launched with the aim of finding process variable settings at which the specified objectives would be met.

### 20.5.2 Experimental Strategy and Results

#### Planning

With the response variables identified as  $y_1$ , Yield (%), and  $y_2$ , Adhesion (grams), a thorough consideration of all aspects of the process led to the following list of seven potential factors—*independent process variables* that could potentially affect yield and adhesion:

1. Amount of additive
2. Raw material supplier
3. Reactor configuration
4. Reactor level
5. Reactor pressure
6. Reactor temperature
7. Reaction Time

As a result, the following overall strategy was devised: first, a set of *screening experiments* will be performed to determine which of these seven variables are important factors; next, a set of *optimization studies* will be carried out to determine the best settings for the important factors; and finally, the optimum setting will be verified in a set of *confirmation experiments*.

These considerations led to the choice of a  $2^{7-3}$  fractional factorial design for the screening, followed by a response surface design for optimization.

<sup>3</sup> Adapted from an example used in an E.I. duPont de Nemours and Company’s Central Research & Development training course; the original source is unknown.

**TABLE 20.8:** Response surface design and experimental results for coating process

| RunOrder | Additive | Time | Temperature | Yield | Adhesion |
|----------|----------|------|-------------|-------|----------|
| 1        | 0        | 20   | 100         | 68    | 3        |
| 2        | 70       | 20   | 100         | 51    | 40       |
| 3        | 35       | 40   | 100         | 75    | 31       |
| 4        | 0        | 60   | 100         | 81    | 10       |
| 5        | 70       | 60   | 100         | 65    | 48       |
| 6        | 35       | 20   | 140         | 80    | 38       |
| 7        | 0        | 40   | 140         | 68    | 24       |
| 8        | 35       | 40   | 140         | 82    | 37       |
| 9        | 35       | 40   | 140         | 87    | 41       |
| 10       | 35       | 40   | 140         | 87    | 40       |
| 11       | 35       | 40   | 140         | 82    | 40       |
| 12       | 35       | 40   | 140         | 85    | 42       |
| 13       | 35       | 40   | 140         | 85    | 42       |
| 14       | 70       | 40   | 140         | 75    | 44       |
| 15       | 35       | 60   | 140         | 92    | 41       |
| 16       | 0        | 20   | 180         | 40    | 37       |
| 17       | 70       | 20   | 180         | 75    | 31       |
| 18       | 35       | 40   | 180         | 77    | 44       |
| 19       | 0        | 60   | 180         | 50    | 40       |
| 20       | 70       | 60   | 180         | 90    | 39       |

### Design and Implementation

The results of the fractional factorial experiments (not shown) identified the following three significant factors, along with appropriate low and high settings:

1.  $x_1$ : Amount of additive; (0, 70);
2.  $x_2$ : Reaction time; (20, 60) mins;
3.  $x_3$ : Reactor temperature; (100, 180) °C.

A three-factor face-centered cube response surface design was therefore used for the optimization experiments. The design (the standard design of 17 runs plus an additional set of center point replicates) and the experimental results are shown in Table 20.8.

#### 20.5.3 Analysis

With the design and data in a MINITAB worksheet, the data analysis is carried out with the sequence: Stat > DOE > Response Surface >

Analyze Response Surface Design > which opens a self-explanatory dialog box. Upon selecting the appropriate options, the following results are obtained, first for Yield:

#### Response Surface Regression: Yield versus Additive, Time, Temperature

The analysis was done using coded units.

##### Estimated Regression Coefficients for Yield

| Term                    | Coef     | SE Coef | T       | P     |
|-------------------------|----------|---------|---------|-------|
| Constant                | 84.5455  | 0.6964  | 121.403 | 0.000 |
| Additive                | 4.9000   | 0.6406  | 7.649   | 0.000 |
| Time                    | 6.4000   | 0.6406  | 9.991   | 0.000 |
| Temperature             | -0.8000  | 0.6406  | -1.249  | 0.240 |
| Additive*Additive       | -12.8636 | 1.2216  | -10.530 | 0.000 |
| Time*Time               | 1.6364   | 1.2216  | 1.340   | 0.210 |
| Temperature*Temperature | -8.3636  | 1.2216  | -6.847  | 0.000 |
| Additive*Time           | 0.7500   | 0.7162  | 1.047   | 0.320 |
| Additive*Temperature    | 13.5000  | 0.7162  | 18.849  | 0.000 |
| Time*Temperature        | -0.2500  | 0.7162  | -0.349  | 0.734 |

S = 2.02574 PRESS = 184.619

R-Sq = 98.92% R-Sq(pred) = 95.13% R-Sq(adj) = 97.94%

##### Estimated Regression Coefficients for Yield using data in uncoded units

| Term                    | Coef         |
|-------------------------|--------------|
| Constant                | 7.87273      |
| Additive                | -0.517792    |
| Time                    | -0.00102273  |
| Temperature             | 1.11864      |
| Additive*Additive       | -0.0105009   |
| Time*Time               | 0.00409091   |
| Temperature*Temperature | -0.00522727  |
| Additive*Time           | 0.00107143   |
| Additive*Temperature    | 0.00964286   |
| Time*Temperature        | -3.12500E-04 |

(The ANOVA table—not shown—displays the typical break down of the sources of variability and establishes that the composite linear, square and interaction terms are all significant.)

The corresponding results for Adhesion are as follows:

#### Response Surface Regression: Adhesion versus Additive, Time, Temperature

The analysis was done using coded units.

**Estimated Regression Coefficients for Adhesion**

| Term                    | Coef     | SE Coef | T       | P     |
|-------------------------|----------|---------|---------|-------|
| Constant                | 40.2364  | 0.5847  | 68.816  | 0.000 |
| Additive                | 8.8000   | 0.5378  | 16.362  | 0.000 |
| Time                    | 2.9000   | 0.5378  | 5.392   | 0.000 |
| Temperature             | 5.9000   | 0.5378  | 10.970  | 0.000 |
| Additive*Additive       | -6.0909  | 1.0256  | -5.939  | 0.000 |
| Time*Time               | -0.5909  | 1.0256  | -0.576  | 0.577 |
| Temperature*Temperature | -2.5909  | 1.0256  | -2.526  | 0.030 |
| Additive*Time           | 0.7500   | 0.6013  | 1.247   | 0.241 |
| Additive*Temperature    | -10.2500 | 0.6013  | -17.046 | 0.000 |
| Time*Temperature        | -0.5000  | 0.6013  | -0.831  | 0.425 |

S = 1.70080 PRESS = 142.826

R-Sq = 98.81% R-Sq(pred) = 94.12% R-Sq(adj) = 97.74%

**Estimated Regression Coefficients for Adhesion using data in uncoded units**

| Term                    | Coef         |
|-------------------------|--------------|
| Constant                | -73.0818     |
| Additive                | 1.58162      |
| Time                    | 0.313182     |
| Temperature             | 0.882159     |
| Additive*Additive       | -0.00497217  |
| Time*Time               | -0.00147727  |
| Temperature*Temperature | -0.00161932  |
| Additive*Time           | 0.00107143   |
| Additive*Temperature    | -0.00732143  |
| Time*Temperature        | -6.25000E-04 |

(The ANOVA table is not shown, neither is the diagnostic warning about an unusual observation, Observation 8, with a standardized residual of -2.03. This value is not so high as to cause serious concern.)

Upon eliminating the coefficients with *p*-values greater than 0.05, we obtain the following response surface models, for Yield,  $y_1$ ,

$$y_1 = 85.55 + 4.90x_1 + 6.40x_2 - 12.86x_1^2 - 8.36x_3^2 + 13.5x_1x_3 \quad (20.33)$$

and for Adhesion,  $y_2$ :

$$y_2 = 40.23 + 8.80x_1 + 2.90x_2 + 5.9x_3 - 6.09x_1^2 - 2.59x_3^2 - 10.25x_1x_3 \quad (20.34)$$

in terms of coded units,

$$x_1 = \frac{A - 35}{35} \quad (20.35)$$

$$x_2 = \frac{\tau - 40}{20} \quad (20.36)$$

$$x_3 = \frac{T - 140}{40} \quad (20.37)$$

where  $A$  =Additive,  $\tau$  =Time, and  $T$  =Temperature, in the original units.

In terms of these original uncoded units, the response surface model equations are:

$$y_1 = 7.873 - 0.518A - 0.001\tau - 0.11A^2 - 0.005T^2 + 0.010AT \quad (20.38)$$

for Yield, and

$$y_2 = -73.082 + 1.582A + 0.313\tau + 0.884T - 0.005A^2 - 0.002T^2 - 0.007AT \quad (20.39)$$

for Adhesion. A plot of the standardized residuals versus fit, and a normal probability plot of the standardized residuals are shown for the Yield and Adhesion responses in Figs 20.9 and 20.10, respectively; neither set of plots shows anything that will invalidate the normality assumptions. This model may now be used to optimize the process.

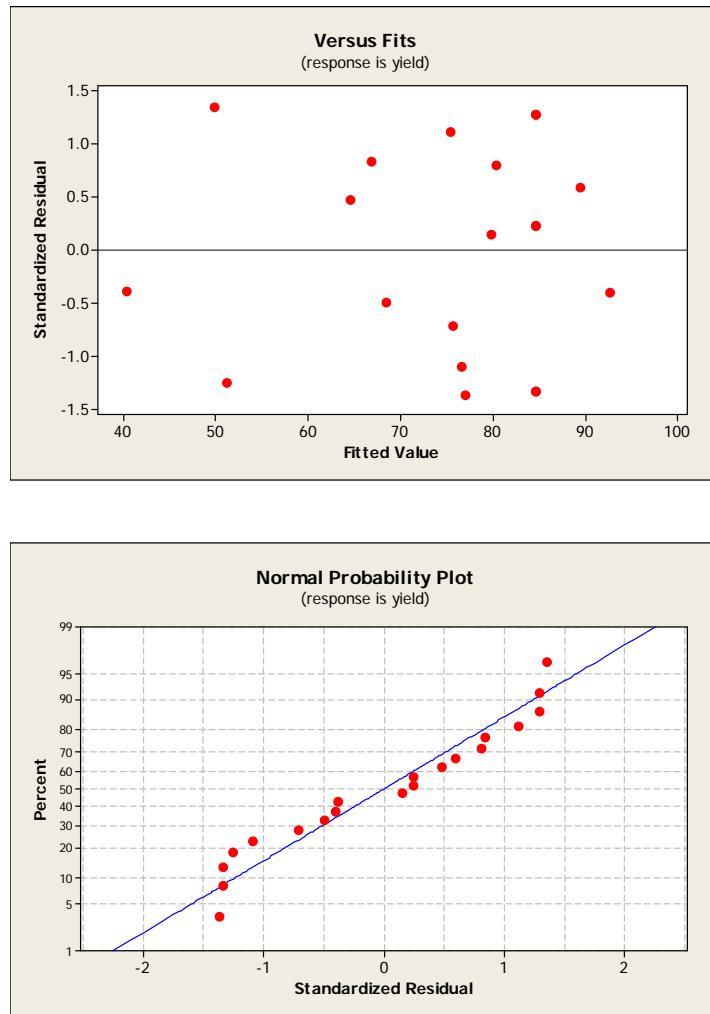
### Optimization

Before engaging in any rigorous analysis, a close examination of these model equations reveals several noteworthy features:

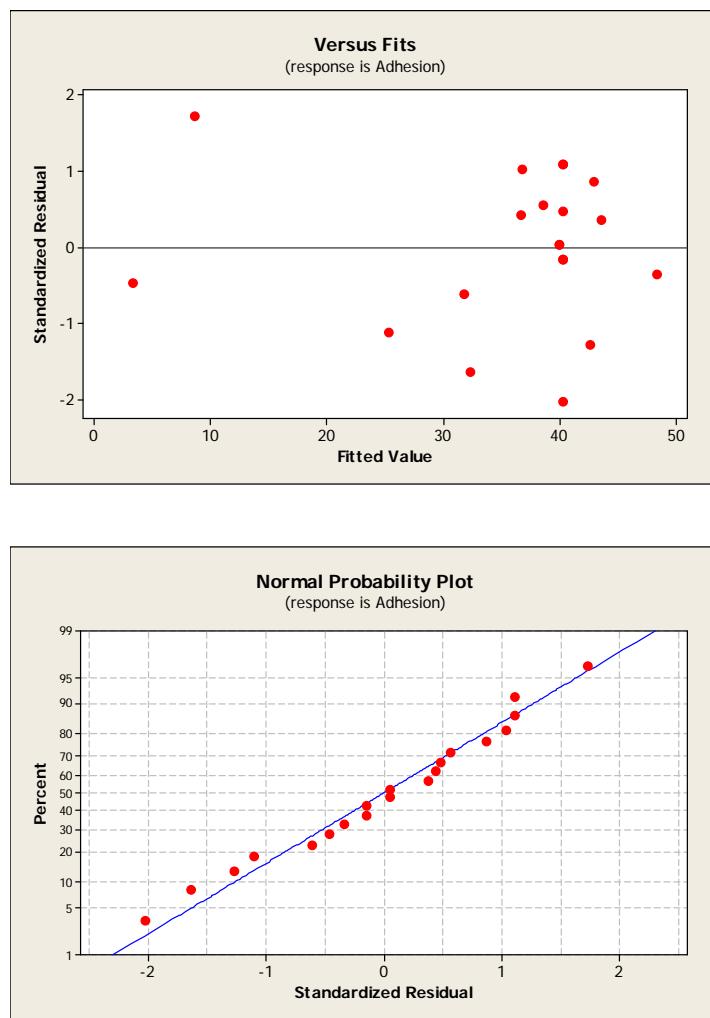
1. The scaled time variable,  $x_2$ , appears in each model as a single, linear term, with a positive coefficient. This implies that each response is linearly dependent on  $x_2$  in a monotonically increasing fashion. As a result, both  $y_1$  and  $y_2$  are maximized with respect to  $x_2$  at the maximum allowable setting.
2. The dependence of  $y_1$  on  $x_1$  is through the pure linear and quadratic terms, and the bilinear  $x_1x_3$  interaction term. The dependence of  $y_1$  on  $x_3$ , on the other hand, is through the quadratic term and bilinear  $x_1x_3$  term; there is no separate linear dependence on  $x_3$ .
3. The dependence of  $y_2$  on  $x_1$  and  $x_3$  is through linear and quadratic terms in each variable as well as through the bilinear  $x_1x_3$  term.

These model characteristics may be visualized graphically a number of different ways, including the popular surface and contour plots. The surface plot is a 3-dimensional plot of the predicted model response,  $\hat{y}$ , as a function of two of the independent variables. The contour plot, on the other hand, consists of a two-dimensional collection of lines of equal values of  $y$ , as a function of two independent variables. (The contour plot is essentially a projection of the surface plot onto the “floor” of the two-dimensional independent variable plane.)

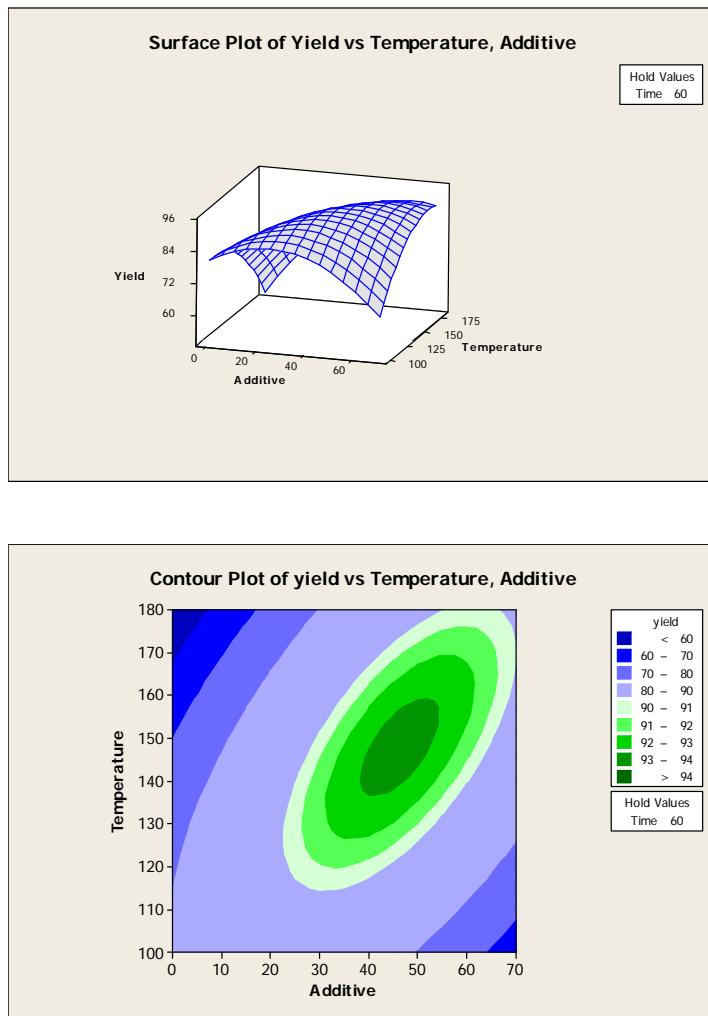
Such a combination of surface and contour plots for Yield and Adhesion as functions of Additive and Temperature (holding Time at the high value of 60) are shown in Figs 20.11 and 20.12. These figures show visually how each response behaves as a function of the factors; the figures also show where the



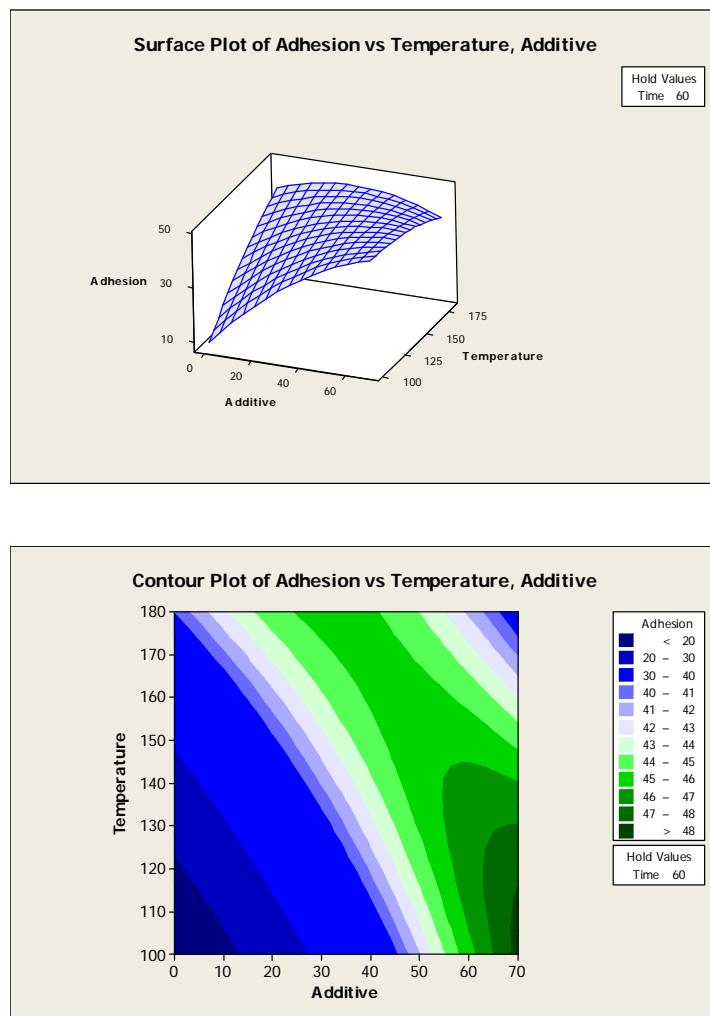
**FIGURE 20.9:** Standardized residual plots for “Yield” response surface model: versus fitted value, and normal probability plot.



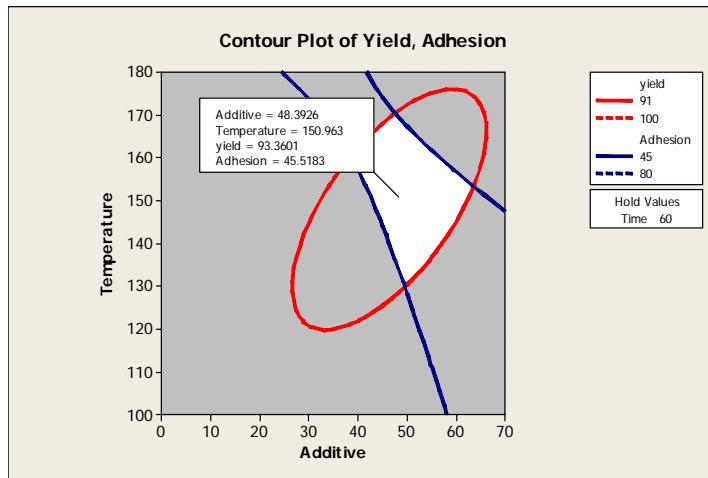
**FIGURE 20.10:** Standardized residual plots for “Adhesion” response surface model: versus fitted value, and normal probability plot.



**FIGURE 20.11:** Response surface and contour plots for "Yield" as a function of Additive and Temperature (with Time held at 60.00).



**FIGURE 20.12:** Response surface and contour plots for “Adhesion” as a function of Additive and Temperature (with Time held at 60.00).



**FIGURE 20.13:** Overlaid contours for “Yield” and “Adhesion” showing feasible region for desired optimum. The planted “flag” indicates the optimum values of the responses along with the corresponding setting of the factors Additive and Temperature (with Time held at 60.00) that achieve this optimum.

optimum might lie. Observe that while the yield response shows the existence of a maximum, the adhesion response shows a saddle point.

At this point, several options are available for determining the optimum settings for these factors: the calculus method, by taking derivatives in Eqs (20.33) and (20.34), (after setting  $x_2$  to its maximum value of 1) and solving simultaneously for  $x_1$ , and  $x_3$ , subject to the constraints in the objectives; or by using various graphically based options available in MINITAB. Since two different responses are involved, to take advantage of the MINITAB options, it is necessary to overlay the two contours for Yield and Adhesion to see the region where the objectives are met simultaneously. The MINITAB contour overlay option, when given the desired values of yield,  $91 < y_1 < 100$ , and desired values for adhesion,  $45 < y_2 < 80$ , (the value of 80 is simply a high enough upper limit) produces the overlaid contour plots shown in Fig 20.13; it indicates the feasible region as the intersection of the two contours. MINITAB has a built-in response optimizer that can be used to find the optimum; it also has a “plant-the-flag” option that allows the user to roam over the feasible region in the overlaid contour plot with the computer mouse, while the values of the responses at the particular location visited literally scroll by on the screen. This is another option for finding the optimum.

While the reader is encouraged to explore all the other options, what we show here in Fig 20.13 is the MINITAB flag planted at the optimum values found by this exploratory “plant-the-flag” option. The optimum responses are:

$$y_1^* = 93.36\%; y_2^* = 45.52 \quad (20.40)$$

found for the factor settings:

$$\text{Additive} = 48.39; \text{Time} = 60; \text{Temperature} = 151.00 \quad (20.41)$$

### Confirmation

A final confirmation set of 5 experimental runs, four  $2^2$  factorial experiments run in a small region around these optimum settings,  $46 < \text{Additive} < 50$ ;  $140 < \text{Temperature} < 160$ , (with Time set at 60 mins), plus one at the prescribed optimum itself, resulted in products all with acceptable yield and adhesions. (Results not shown).

---

## 20.6 Summary and Conclusions

The basic premise of Chapters 12–19 is that whenever variability and uncertainty are intrinsic to a problem, statistics, building on probability, provides a consistent set of tools for handling such problems systematically. However, using simple, tailor-made, “textbook” examples, to demonstrate how various statistical concepts—sampling, estimation, hypothesis testing, regression, experimental design and analysis—are applied is one thing; solving real-life problems with these statistical techniques is another. Real-life problems are never ideal; also, solving them typically requires choosing the appropriate combination of these tools and using them *appropriately*. This chapter therefore has been devoted to using three classes of real-life problems as a capstone demonstration of statistics in practice. The first category of problems required estimating population parameters from samples, carrying out hypothesis tests about these populations, and using the thus-validated population models to solve non-trivial problems. With the second problem we demonstrated the power of simple regression modeling, and a “forensic” application of hypothesis testing to detect hidden structure in the data; but the problem also served to illustrate that there is significant latitude in carrying out data analysis, since the problem could have been handled several different ways (See Project Assignment 2).

Ironically, the final problem, on the application of design of experiments to industrial process optimization, undoubtedly the most “practical” of the collection, is the one whose structure is closest to “textbook” form: a fractional factorial design to identify significant factors, followed by a response surface design to develop a quadratic response model used for optimization, capped off by factorial confirmation experiments. But the sense of how long it would have taken to solve this problem (if at all) without the systematic experimental design strategy discussed, and how much money (and effort) was saved as a

consequence, impossible to convey adequately in the presentation, are well-appreciated by practitioners of the art.

Taken together then, these case studies illustrate the many faces of real-life problems to which statistics has been successfully applied. The project assignments below are offered as a way to broaden the reader's perspective of the themes illustrated in this chapter.

This chapter joins Chapters 7 and 11 to complete this book's trilogy of chapter-length cases studies; it also concludes Part IV. The remainder of the book, Part V, is devoted to a hand-selected trio of "special topics" each with roots in probability and statistics, but all of which have since evolved into legitimate subject matters in their own rights. These topics are therefore "applications" of probability and statistics but in a much grander sense.

## PROJECT ASSIGNMENTS

### 1. Effect of Bayesian Prior Distributions on Estimation

*Just how much of an effect does the prior distribution used for recursive Bayesian estimation have on the results?* Choose two different gamma pdfs as prior distributions for the unknown Poisson parameter,  $\theta$ , and repeat the recursive Bayesian estimation portion of the Prussian army data analysis case study, using the data in Table 20.3. For each prior distribution,

- Obtain year-by-year estimates; compare them to the results presented in Table 20.4; plot them as in Fig 20.3 with the maximum likelihood estimate.
- Obtain an explicit expression for the final posterior distribution; plot the prior and the final posterior distributions as in Fig 20.4.

Write a report on your analysis, discussing the effects of the prior distributions you chose on the recursive parameter estimation process.

### 2. "First Principles" Population Dynamics Modeling

Consult references on the mathematical modeling of biological populations (e.g., Brauer and Castillo-Chavez, (2001)<sup>4</sup>), and use these concepts to develop an alternative model to represent the US Population data in Table 20.6. The following two things are required:

- Use only data up to and including 1970 to develop the model; validate the model by using it to predict the 1980, 1990 and 2000 census results.
- Use your validated model to predict the 2010 census.

---

<sup>4</sup>Brauer, F. and C. Castillo-Chavez (2001) . *Mathematical Models in Population Biology and Epidemiology*, Springer-Verlag, NY.

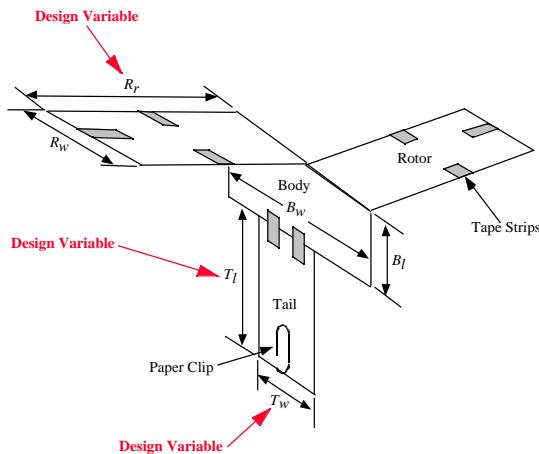


FIGURE 20.14: Schematic diagram of folded helicopter prototype

Write a report on your model development, validation and model prediction. You have considerable latitude in how to complete this assignment. Be creative.

### 3. Experimental Design and Analysis

The *length differential* between the ring and index fingers on the human hand has been postulated to be a subtle but not yet well-understood distinguishing factor between men and women. You are required to find  $n_M$  male subjects and  $n_F$  female subjects, acquire and analyze data on the ring-index fingers length differential,  $\delta_{RI}$ , and confirm or refute this postulate.

Write a report in which you state the project objectives, justify your choice of sample sizes  $n_M$  and  $n_F$ , describe your measurement procedure clearly, show your data tables (include the first names of your subjects), and present your analysis of the data and your conclusions.

### 4. Process Development and Optimization

The objective of this assignment is to develop a paper helicopter that has the maximum “hang time” from a fixed height, and to develop a model that will allow you to predict the hang time based on whatever design features you find to be important. Apply a comprehensive design of experiments strategy.

A schematic diagram of the folded helicopter is shown in Fig 20.14; a beginning “template” that can be photocopied unto blank sheets of paper and then cut and folded to make the various prototypes, is shown in Fig 20.15.

- Consider beginning with screening experiments to identify the factors that affect the helicopter’s flight time.

- Conduct a series of experiments that will ultimately lead to a mathematical model and an “optimal” design.
- Predict the maximum flight time and perform experiments to confirm this prediction.

Write a report summarizing your design at the prototype stage, and the analysis leading to the optimum design. Discuss your analysis methods and show the final design, the results of the model predictions, and the confirmation of your predictions.

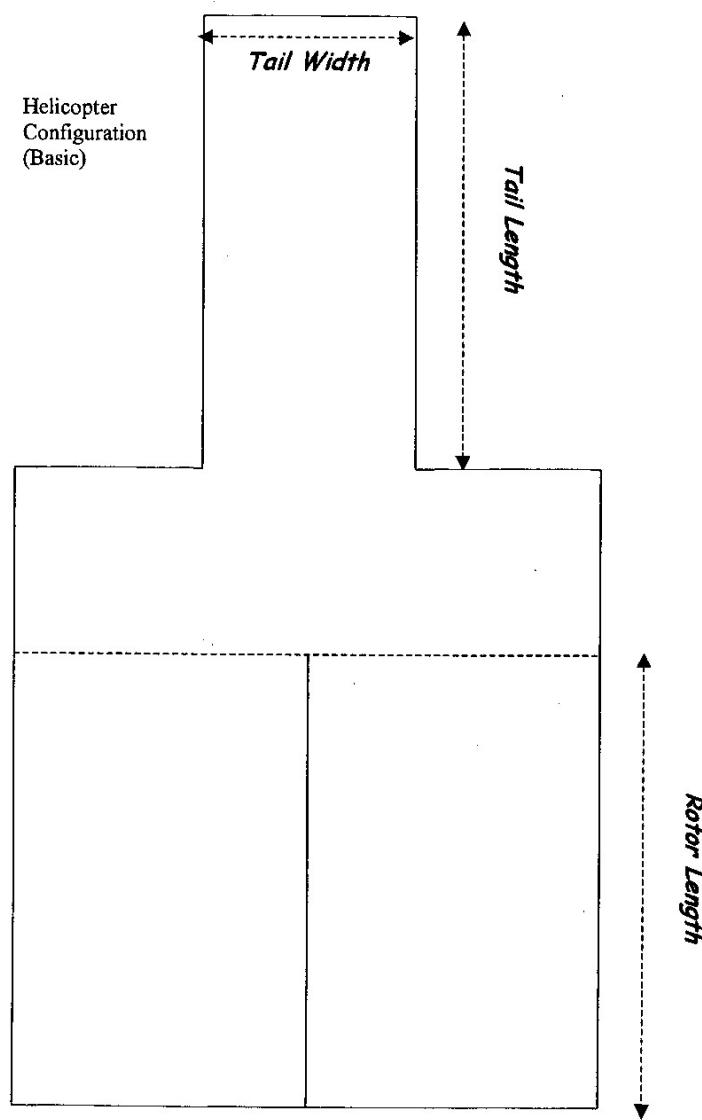


FIGURE 20.15: Paper helicopter prototype



Part V

# Applications

—

|

|

—

—

|

|

—

## Part V: Applications

*Dealing with Random Variability in Practice*

---

*What is laid down, ordered, factual, is never enough to embrace  
the whole truth; life always spills over the rim of every cup.*

Boris Pasternak (1890–1960)

# Part V: Applications

*Dealing with Random Variability in Practice*

- **Chapter 21:** Reliability and Life Testing
- **Chapter 22:** Quality Assurance and Control
- **Chapter 23:** Introduction to Multivariate Analysis

# Chapter 21

## Reliability and Life Testing

|        |  |     |
|--------|--|-----|
| 21.1   | Introduction .....   | 900 |
| 21.2   | System Reliability .....                                   | 901 |
| 21.2.1 | Simple Systems .....                                       | 901 |
|        | Series Systems .....                                       | 901 |
|        | Parallel Systems .....                                     | 903 |
|        | Components and Modules .....                               | 905 |
| 21.2.2 | Complex Systems .....                                      | 906 |
|        | Series-Parallel Configurations .....                       | 906 |
|        | $k$ -of- $n$ Parallel System .....                         | 907 |
|        | Systems with Cross-links .....                             | 909 |
| 21.3   | System Lifetime and Failure-Time Distributions .....       | 910 |
| 21.3.1 | Characterizing Time-to-Failure .....                       | 911 |
|        | The Survival Function, $S(t)$ .....                        | 911 |
|        | The Hazard Function, $h(t)$ : .....                        | 911 |
| 21.3.2 | Probability Models for Distribution of Failure Times ..... | 913 |
| 21.4   | The Exponential Reliability Model .....                    | 914 |
| 21.4.1 | Component Characteristics .....                            | 914 |
| 21.4.2 | Series Configuration .....                                 | 915 |
| 21.4.3 | Parallel Configuration .....                               | 916 |
| 21.4.4 | $m$ -of- $n$ Parallel Systems .....                        | 917 |
| 21.5   | The Weibull Reliability Model .....                        | 917 |
| 21.6   | Life Testing .....   | 918 |
| 21.6.1 | The Exponential Model .....                                | 919 |
|        | Estimation .....   | 920 |
|        | Precision of Estimates .....                               | 920 |
|        | Hypothesis Testing .....                                   | 920 |
| 21.6.2 | The Weibull Model .....                                    | 922 |
| 21.7   | Summary and Conclusions .....                              | 922 |
|        | REVIEW QUESTIONS .....                                     | 923 |
|        | EXERCISES AND APPLICATION PROBLEMS .....                   | 925 |

*The shifts of Fortune test the reliability of friends.*

Cicero (106–43 BC)

The aphorism “Nothing lasts forever” (or any of its sundry variations), has long served philosophers, ancient and modern, as a concise way to convey the transient and ephemeral nature of the world around us. Specifically for the material things, however, the issue is never so much about *not* lasting forever, which is certain; it is more about *how long* they will last, which is uncertain. These manufactured and engineered products consist of components with finite functioning lifetimes, and the failure of any of the individual constituent

components has repercussions on the ability of the overall system to function. But the failure of a component, or of the overall system, is subject to random variability, so that items manufactured at the same site by the same crew of operators and used essentially in the same manner will fail at different times.

*Reliability*, the attribute of an individual component or a system which indicates how long it is expected to function as prescribed, will be studied in this chapter. With roots deeply anchored in probability theory and statistics, reliability theory, and life testing, its experimental counterpart, have jointly evolved into an important and extensive field of study. The discussion here is therefore designed to be illustrative rather than exhaustive. Still, enough of the essential material will be presented to provide the reader with an appreciation of the basic principles and practical applications.

---

## 21.1 Introduction

Engineered and natural systems—chemical processes, mechanical equipment, electrical devices, even the human body, etc.,—consist of individual units connected in a logical fashion for achieving specific overall system goals. A basic principle underlying such systems is that how the overall system performs depends on the individual component's performance and how the components are connected. Plant equipment do fail, as do mechanical and electrical systems; automobiles break down; and human beings fall sick and eventually die. While the issue of safety and the *consequences* of system failure typically dominate most discussions about system performance, of equal importance is the issue of how “reliable” these various systems and their constituent components are. For how long will the car start every morning? How long can the entire refinery operate before we need to shut it down for maintenance? How long will the new dishwasher last? These are all issues of *reliability*, a concept deeply influenced by variability and uncertainty surrounding such questions. The subject matter of reliability therefore relies heavily on probability theory and statistics. The following is one version of a formal definition:

The reliability of a component, or a system (of components), is the probability that it will function properly within *prescribed* limits for at least a *specified* period of time, under *specified* operating conditions.

All the qualifiers italicized in this definition are important. For example, what is satisfactory for laboratory use may be inadequate for the harsh commercial

plant environment; and what is expected of an inexpensive disposable pen is different from what is expected of the more expensive brand. Giving all due respect to such qualifiers, we may now give the following mathematical definition:

The reliability of a component or a system,  $R(t)$ , is the probability that it “survives” for a specified period of time,  $t$ , i.e.,

$$R(t) = P(T > t) \quad (21.1)$$

“Component reliability,”  $R_i(t)$ , is the reliability of an individual component,  $i$ ; while “system reliability,”  $R_s(t)$ , is the reliability of the overall system. Clearly,  $R_s$  is a function of  $R_i$ , and given  $R_i$  and how the components are connected to constitute the overall system, one can compute  $R_s$  using techniques discussed in the next section.

## 21.2 System Reliability

As one would expect, the overall reliability of a system is determined by:

1.  $R_i$ , the reliability of constituent components; and,
2. How the components are connected to form the overall system, i.e., the system configuration.

The defining problem of system reliability is therefore easily stated: Given  $R_i$  and the system configuration, find  $R_s$ . The presentation in this section is therefore organized around system configurations, beginning with “simple” systems and building up to more “complex” ones. The companion issue of how the individual reliabilities are determined belongs under the topic of statistical “life testing,” which is discussed in Section 21.6.

### 21.2.1 Simple Systems

The  $n$  components  $C_i : i = 1, 2, \dots, n$ , of a system can be configured in series, or in parallel, as shown in Fig 21.1, or as a combination series-parallel arrangement, as in Fig 21.2 (shown here for  $n = 6$ ), with or without cross-linking between the components. A system is said to be “simple” if it consists of either a straightforward series arrangement, or a straightforward parallel arrangement; otherwise it is said to be “complex.”

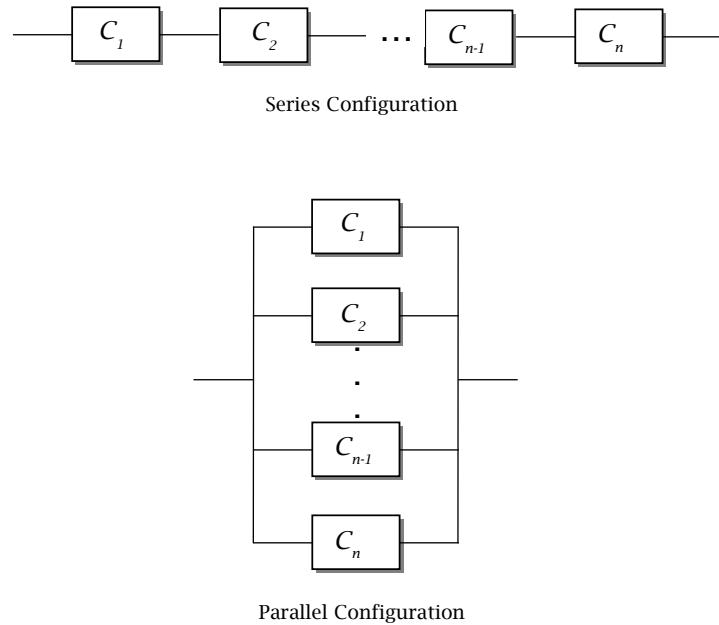


FIGURE 21.1: Simple Systems: Series and parallel configuration

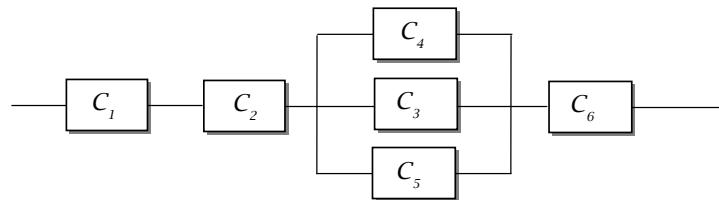


FIGURE 21.2: A series-parallel arrangement of a 6-component system

### Series Systems

Consider the system depicted in the top panel of Fig 21.1, where the components are connected in series and the reliability of component  $C_i$  is  $R_i$ . If each component operates mutually independently of the others, by which we mean that the performance of one component has no effect on the performance of any other component, then:

1. If any component fails, the entire system fails; consequently,
2. The system reliability is obtained as

$$R_s = \prod_{i=1}^n R_i \quad (21.2)$$

Eq (21.2) arises because, by definition,  $R_s$  is the probability that the entire system functions, i.e.,  $P(C_1 \text{ and } C_2 \text{ and } \dots \text{ and } C_n \text{ all function})$ . By independence, therefore,

$$\begin{aligned} R_s &= P(C_1 \text{ functions}) \times P(C_2 \text{ functions}) \times \dots \times P(C_n \text{ functions}) \\ &= R_1 R_2 \cdots R_n \end{aligned} \quad (21.3)$$

as required. Eq (21.2) is known as the *product law of reliabilities*. Now, because reliability is a probability, it follows that  $0 < R_i < 1$ ; as a result, an important implication of this product law of reliabilities is that as  $n$  increases,  $R_s$  decreases. Thus, a system's reliability decreases as the number of components increases. Intuitively, this makes sense: as we increase the number of components that *must* function simultaneously for the entire system to function, we provide more opportunities for something to go wrong, thereby decreasing the probability that the entire system will function.

#### **Example 21.1: RELIABILITY OF 4-COMPONENT AND 6-COMPONENT SERIES SYSTEMS**

If four identical components each with reliability  $R_i = 0.98$  are connected in series, the system reliability is

$$R_{s4} = 0.98^4 = 0.922 \quad (21.4)$$

If two more identical components are added in series, the system reliability becomes

$$R_{s6} = 0.922 \times 0.98^2 = 0.886 \quad (21.5)$$

The system with the higher number of components in series is therefore seen to be much less reliable.

Note that as a consequence of the basic laws of the arithmetic operation of multiplication, system reliability for a series configuration is independent of the order in which the components are arranged.

### Parallel Systems

Now consider the system depicted in the bottom panel of Fig 21.1 where the components are arranged in parallel, and again, the reliability of component  $C_i$  is  $R_i$ . Observe that in this case, if one component fails, the entire system does not necessarily fail. In the simplest case, the system fails when all  $n$  components fail. In the special “ $k$ -of- $n$ ” case, the system will function if at least  $k$  of the  $n$  components function. Let us consider the simpler case first.

In the case when the system fails only if all  $n$  components fail, then  $R_s$  is the probability that *at least one* component functions, which is equivalent to  $1 - P(\text{no component functions})$ . Now, let  $F_i$  be the “unreliability” of component  $i$ , the probability that the component does *not* function; then by definition,

$$F_i = 1 - R_i \quad (21.6)$$

If  $F_s$  is the system “unreliability,” i.e., the probability that *no* component in the system functions, then by independence,

$$F_s = \prod_{i=1}^n F_i \quad (21.7)$$

and since  $R_s = 1 - F_s$ , we obtain

$$R_s = 1 - \prod_{i=1}^n (1 - R_i) \quad (21.8)$$

For parallel systems, therefore, we have the *product law of unreliabilities* expressed in Eq (21.7), from which Eq (21.8) follows. Specific cases of Eq (21.8) can be informative, as the next example illustrates.

#### **Example 21.2: RELIABILITY OF 2-COMPONENT PARALLEL SYSTEM**

Obtain an explicit expression for the reliability of a system consisting of a parallel arrangement of two components,  $C_1$  and  $C_2$ , with respective reliabilities,  $R_1$  and  $R_2$ . Explain in words what the expression for the system reliability means in terms of the status of each component.

#### **Solution:**

From Eq (21.8), we have, for the two component system,

$$R_s = 1 - (1 - R_1)(1 - R_2) = R_1 + R_2 - R_1 R_2 \quad (21.9)$$

This expression can be rearranged in one of two equivalent ways:

$$R_s = R_1 + R_2(1 - R_1) = R_1 + R_2 F_1 \quad (21.10)$$

$$R_s = R_2 + R_1(1 - R_2) = R_2 + R_1 F_2 \quad (21.11)$$

In words, Eq (21.10) indicates that the entire system functions if (a)

$C_1$  functions regardless of the status of component  $C_2$  (with a probability  $R_1$ ), or (b)  $C_2$  functions when  $C_1$  fails, with probability  $R_2F_1$ . Eq (21.11) expresses the mirror image circumstance. Thus, these two equivalent expressions show how, in this parallel arrangement, one component serves as a backup for the other.

Since  $0 < R_i < 1$ , it is also true that  $0 < (1 - R_i) < 1$ . As a result of Eq (21.8), as  $n$  increases in a parallel configuration,  $R_s$  also increases. Again, this makes sense intuitively: each additional component in the parallel configuration provides an additional level of redundancy, thereby increasing the probability that at least one of the components will continue to function.

**Example 21.3: RELIABILITY OF 2-COMPONENT AND 4-COMPONENT PARALLEL SYSTEMS**

If two of the identical components of Example 21.1 are now arranged in parallel (instead of in series), first obtain the system reliability for this 2-component parallel configuration. If two more identical components are added in parallel, obtain the system reliability for the resulting 4-component parallel configuration.

**Solution:**

In this case, the system reliability for the 2-component system is

$$R_s = 1 - (1 - 0.98)^2 = 1 - (0.02)^2 = 0.9996 \quad (21.12)$$

When two more components are added in parallel, the system reliability becomes

$$R_s = 1 - (0.02)^4 = 0.99999984 \quad (21.13)$$

where it is necessary to retain so many decimal places to see that the system does not quite possess absolutely perfect reliability, but it is very close.

Thus, we see that by adding more components in parallel, the system reliability is improved substantially, a reverse of the case with the series arrangement.

Again, from Eq (21.8) and the laws of multiplication, the order in which the components are arranged in the parallel configuration is immaterial to the value of  $R_s$ .

### Components and Modules

If a box drawn around a set of component blocks in the system's representation has a single line going into the box, and a single line coming out of it, the collection of components is a called a *module*. For example, the entire collection of components in the series arrangement in Fig 21.1 constitutes a module. Of course, several smaller modules can be created from this larger one by drawing the module box to contain fewer components. A single component is the smallest (simplest) module. Observe that the entire collection of

components in the parallel arrangement in Fig 21.1 also constitutes a module, but smaller modules can also be created from this largest one.

The reliability of a system consisting entirely of modules is obtained by first finding the module reliabilities and then combining these according to how the modules are configured to constitute the complete system. For simple systems, the individual components are all modules, which is why reliability analysis for such systems can proceed in the direct manner discussed above. This is not so for complex systems.

### 21.2.2 Complex Systems

Complex systems arise from a combination of simple systems. This fact makes it possible to invoke the results and concepts developed earlier in obtaining the reliabilities of complex systems. How these results are employed depends on the nature of the “complexity” — the configuration and the requirements, as we now discuss.

#### Series-Parallel Configurations

Series-parallel systems are complex because they consist of a mixture of simple series and simple parallel systems. The analysis technique therefore consists of first consolidating each parallel ensemble in the system into a single module; this has the net effect of reducing the system to one containing only series modules (since the original components in the series arrangement are all modules) so that the results obtained earlier can then be applied.

Let us use the system in Fig 21.2 to illustrate. We begin by consolidating the parallel subsystem consisting of components  $C_3$ ,  $C_4$ , and  $C_5$  into a single composite module, say  $C_{345}$ , with reliability  $R_{345}$ . Then, from Eq (21.8),

$$R_{345} = 1 - (1 - R_3)(1 - R_4)(1 - R_5) \quad (21.14)$$

As a result of this consolidation, the entire system now consists strictly of a series arrangement of three single components,  $C_1$ ,  $C_2$  and  $C_6$ , along with the composite module  $C_{345}$ , with the result that the system reliability is:

$$R_s = R_1 R_2 R_{345} R_6 \quad (21.15)$$

$$= R_1 R_2 R_6 [1 - (1 - R_3)(1 - R_4)(1 - R_5)] \quad (21.16)$$

The following example illustrates the application of these principles.

#### **Example 21.4: RELIABILITY OF SAMPLING-ANALYZER SYSTEM**

A system for analyzing the composition of a laboratory-scale distillation column’s products consists of a side-stream sampling pump, a solenoid valve and a densitometer. (1) Obtain the overall reliability of the sampling-analyzer system if the components are configured as in Fig 21.3, with the reliability of each component as shown. (2) Because

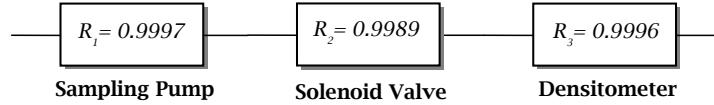


FIGURE 21.3: Sampling-analyzer system: basic configuration

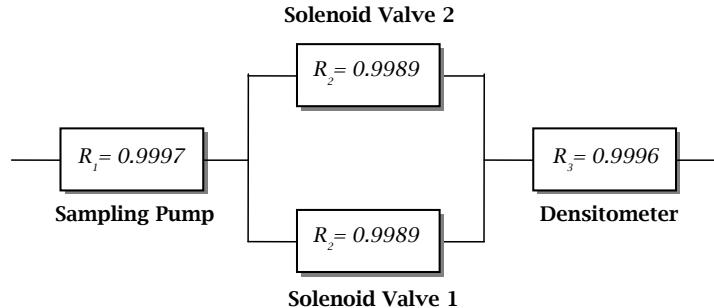


FIGURE 21.4: Sampling-analyzer system: configuration with redundant solenoid valve

the solenoid valve is the “least reliable” of the components, it has been suggested to add another identical solenoid valve in parallel for redundancy, resulting in the configuration in Fig 21.4. Obtain the system reliability for this new configuration and compare it to that for the basic configuration.

**Solution:**

- (1) The system reliability for this simple series configuration is obtained as a product of the indicated reliabilities, or

$$R_s = 0.9982 \quad (21.17)$$

indicating a 99.82% system reliability, or a 0.18% chance of system failure.

- (2) By first consolidating the parallel solenoid valve components, the reliability of the complex system is obtained as:

$$R_s = 0.9997 \times [1 - (1 - 0.9989)^2] \times 0.9996 = 0.9993 \quad (21.18)$$

so that the system reliability is improved from 99.82% to 99.93% by the addition of one more solenoid valve in parallel.

### k-of-n Parallel System

Instead of the simpler case where the system fails if and only if all of the components fail, we now consider the case when at least  $k$  of the  $n$  components are required to function for the entire system to function.

If each component  $C_i$  has identical reliability  $R_i = R$ , then according to the stated requirement,  $R_s$  is the probability that at least  $k$  out of  $n$  components

function. The event that at least  $k$  components out of  $n$  function is composed of several mutually exclusive events:  $E_{k|n}$ , the event that  $k$  components out of  $n$  function and  $n - k$  do not; or  $E_{k+1|n}$ , where  $k + 1$  components function and  $(n - k - 1)$  do not; ... or,  $E_{n|n}$ , all  $n$  function.

Now, because the probability that each  $C_i$  functions is constant and equal to  $R$  (as a result of component independence), observe that the probability that  $k$  components function and  $n - k$  do not function is akin to the binomial problem in which one obtains  $k$  “successes” in  $n$  trials, i.e.,

$$P(E_{k|n}) = \frac{n!}{k!(n-k)!} R^k (1-R)^{n-k} \quad (21.19)$$

Thus, the required system reliability is obtained as:

$$\begin{aligned} R_s &= P(E_{k|n}) + P(E_{k+1|n}) + \cdots + P(E_{n|n}) \\ &= \sum_{i=k}^n \frac{n!}{i!(n-i)!} R^i (1-R)^{n-i} \end{aligned} \quad (21.20)$$

Again, note that because all components  $C_i$  are identical, with identical reliabilities, which of the  $n$  belongs to the functioning group of  $k$  is immaterial.

In the case where

1. The reliabilities might be different, or
2. Specific components  $C_i : i = 1, 2, \dots, (k-1)$  with corresponding reliabilities,  $R_i$ , are required to function, along with at least one of the remaining  $(n - k + 1)$  components,

then, for such a  $k$ -out-of- $n$  system, the system reliability is:

$$R_s = \prod_{i=1}^{k-1} R_i \left[ 1 - \prod_{i=k}^n (1-R_i) \right] \quad (21.21)$$

exactly like  $(k-1)$  components in series connected to  $(n-k+1)$  components in parallel. This is because the stipulation that components  $C_i ; i = 1, 2, \dots, (k-1)$ , function is akin to a series connection, with the remaining  $(n - k + 1)$  as redundant components.

#### **Example 21.5: RELIABILITY OF PARALLEL COMPUTING SYSTEM**

A high-performance computing system used for classified cryptography studies consists of a bank of 10 computers, 7 of which are high-end workstations with equal reliability  $R_i = 0.999$ , configured in parallel; the remaining 3 are backup low-end workstations with equal but lower reliability  $R_i = 0.9$ . The performance requirement is that 6 of the high-end computers *must* function along with *at least one* of the remaining 4, i.e., the workload can be handled only by 7 high-end computers, or by 6 high-end computers plus no more than one low-end backup. What

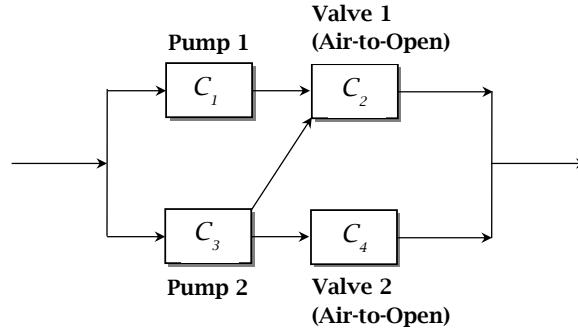


FIGURE 21.5: Fluid flow system with a cross link

is the reliability of this 7-of-10 parallel computing system?

**Solution:**

In this case, from Eq (21.21)

$$\begin{aligned} R_s &= (0.999)^6 \{1 - [(1 - 0.999) \times (1 - 0.9)^3]\} \\ &= 0.9940149 \end{aligned} \quad (21.22)$$

just a hair over 99.4%, the reliability of the module of 6 required high-end workstations. Thus, the combination of the extra high-end workstation plus the 3 back-up low-end ones has the net effect of essentially preserving the reliability of the mandatory module of 6.

### Systems with Cross-links

Consider the system shown in Fig 21.5, a fluid flow system consisting of two pumps and two air-to-open valves with the following failure-mode characteristics: a failed pump is unable to move the fluid through the valve; and the air-to-open valves, which fail shut, are unable to let the fluid through. This system consists of a parallel configuration of two Pump-Valve modules with a cross-link from the lower pump to the upper valve (from component  $C_3$  to  $C_2$ ). Were it not for the presence of this cross-link, the series-parallel arrangement is easily dealt with using results obtained earlier. But as a result of the cross-link, components  $C_3$  and  $C_2$  are not modules, while all the other components are. This complicates matters a bit, so that analyzing systems with cross-links requires special considerations.

First, we need to introduce some useful notation. Let  $P(C_i)$  be the probability that component  $C_i$  functions, and let  $P(C_i^*)$  be the probability that  $C_i$  does not function, i.e., the probability that component  $C_i$  has failed. Note that by definition of reliability,

$$P(C_i) = R_i \quad (21.23)$$

$$P(C_i^*) = 1 - R_i \quad (21.24)$$

Similarly,  $P(S)$ , the probability of the entire system functioning, is  $R_s$ .

To obtain the system reliability,  $R_s$ , for systems with cross-links, we must choose one of the components as the *keystone* component,  $C_k$ , on which the analysis is based. We then compute the following conditional probabilities:  $P(S|C_k)$ , the probability that the system functions given that  $C_k$  is functioning, and  $P(S|C_k^*)$ , the probability that the system functions given that  $C_k$  has failed. From these partial probabilities, we are able to invoke a result from Chapter 3, the “Theorem of total probability” (Eqs (3.47) and (3.53)), to obtain,

$$\begin{aligned} P(S) &= P(S|C_k)P(C_k) + P(S|C_k^*)P(C_k^*) \\ &= P(S|C_k)P(C_k) + P(S|C_k^*)[1 - P(C_k)] \end{aligned} \quad (21.25)$$

Returning to the example in Fig 21.5, let us choose component  $C_3$  as the keystone; i.e.,

$$C_k = C_3 \quad (21.26)$$

We may now observe the following:

- Because  $C_1$  and  $C_3$  are in parallel, if  $C_3$  functions, then the system functions if either  $C_2$  or  $C_4$  functions; the status of  $C_1$  is therefore immaterial in this case. Thus,

$$P(S|C_3) = 1 - (1 - R_2)(1 - R_4) \quad (21.27)$$

- If  $C_3$  does *not* function, the system will function only if  $C_1$  and  $C_2$  function; the status of  $C_4$  is immaterial. And since  $C_1$  and  $C_2$  are in series,

$$P(S|C_3^*) = R_1 R_2 \quad (21.28)$$

And now, from Eq (21.25), we obtain:

$$P(S) = [1 - (1 - R_2)(1 - R_4)]R_3 + R_1 R_2(1 - R_3) \quad (21.29)$$

which simplifies to give:

$$R_s = R_1 R_2 + R_2 R_3 + R_3 R_4 - R_1 R_2 R_3 - R_2 R_3 R_4 \quad (21.30)$$

In principle, it does not matter which component is chosen as the keystone; the same result is obtained. In actual fact, however, the resulting analysis is more straightforward if one of the components associated with the cross-link is chosen as the keystone. As an exercise, the reader should select  $C_1$  as the keystone and use it to obtain the expression for  $R_s$ .

For more complicated systems, one needs as many keystones as there are cross-links.

## 21.3 System Lifetime and Failure-Time Distributions

### 21.3.1 Characterizing Time-to-Failure

From the definition in Eq (21.1), we know that system or component reliability has to do with the probability of the entity in question remaining in service beyond a given time,  $t$ . The so-called “system lifetime” (or component lifetime) is therefore a random variable,  $T$ , having a pdf,  $f(t)$ , sometimes known as the time-to-failure (or failure-time) distribution.

We now recall the discussion in section 4.5 of Chapter 4 and note that even though as a pdf,  $f(t)$  can be studied in its own right, there are other even more relevant ways of characterizing the random variable,  $T$ , in addition to what  $f(t)$  provides. These functions were introduced in Chapter 4, but are re-visited here in their more natural setting.

#### The Survival Function, $S(t)$

The survival function was defined in Chapter 4 as

$$S(t) = P(T > t) = 1 - F(t) \quad (21.31)$$

but, as we know, this is actually identical to the reliability function defined earlier in this chapter, in Eq (21.1). The survival function is therefore also known as the reliability function.

The  $F(t)$  noted above is the cumulative distribution function, which, by definition, is

$$F(t) = \int_0^t f(\tau)d\tau \quad (21.32)$$

But in the specific case where  $f(t)$  is a failure-time distribution, this translates to the probability that a component or system fails before  $T = t$ , making  $F(t)$  the complement of the reliability function (as already implied, of course, by Eq (21.31)). Thus, in lifetime studies, the cdf,  $F(t)$ , is also the same as the system “unreliability,” something we had alluded to earlier in dealing with the parallel system configuration (see Eq (21.6)), but which was presented simply as a definition then.

#### The Hazard Function, $h(t)$ :

This function, defined as:

$$h(t) = \frac{f(t)}{R(t)} = \frac{f(t)}{1 - F(t)} \quad (21.33)$$

is the instantaneous failure rate, or simple “failure rate.” Recall from Chapter 4 that  $h(t)dt$  is the probability of failure in the interval  $(t, t+dt)$ , given survival

until time  $t$ , in precisely the same way that  $f(x)dx$  is the probability of a continuous random variable,  $X$ , taking on values in the interval  $(x, x + dx)$ .

The relationship between the hazard function and several other functions is of some importance in the study of component and system lifetimes. First, it is related to the reliability function as follows:

From the definition of  $R(t)$  as  $1 - F(t)$ , taking first derivatives yields

$$R'(t) = \frac{dR(t)}{dt} = -f(t) \quad (21.34)$$

so that, from Eq (21.33),

$$\begin{aligned} h(t) &= \frac{-R'(t)}{R(t)} \\ &= -\frac{d}{dt}[\ln R(t)] \end{aligned} \quad (21.35)$$

The solution to this ordinary differential equation is easily obtained as:

$$R(t) = e^{-\int_0^t h(\tau)d\tau} \quad (21.36)$$

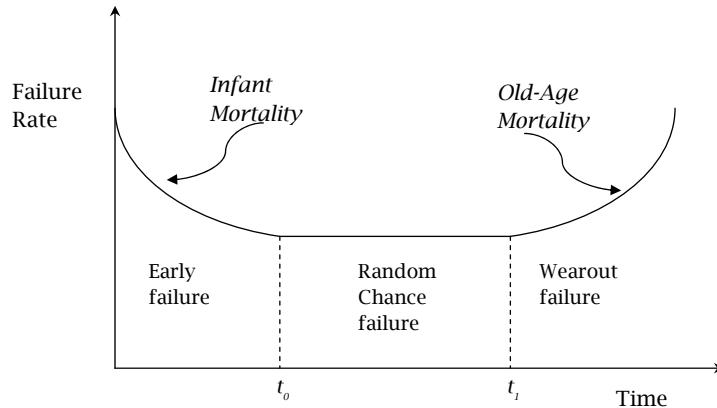
Finally, since, from Eq (21.33)  $f(t) = h(t)R(t)$ , the relationship between the hazard function and the standard pdf  $f(t)$  is

$$f(t) = h(t)e^{-\int_0^t h(\tau)d\tau} \quad (21.37)$$

The typical hazard function (or equivalently, failure rate) curve is shown in Fig 21.6. This is the so-called “bathtub” curve for representing failure characteristics of many realistic systems, including human mortality. Before discussing the characteristics of this curve, it is important, first, to clear up a popular misconception. The “rate” in failure rate is not with respect to time; rather it is the proportion (or percentage) of the components surviving until time  $t$  that are expected to fail in the infinitesimal interval  $(t, t + \Delta t)$ . This “rate” is comparable to the “rate” in interest rate in finance, which refers not to time, but to the proportion of principal borrowed.

The failure rate curve is characterized by 3 distinct parts:

1. *Initial Period*:  $t \leq t_0$ , characterized by a relatively high failure rate that decreases as a function of time. This is the “early failure” period where inferior items in the population fail quickly. This is also known as the “infant mortality” period, the analogous characteristic in human populations.
2. “Normal” Period:  $t_0 \leq t \leq t_1$ , characterized by constant failure rate. This is the period of useful life of many products where failure is due to purely random chance, not systematic problems.
3. *Final period*:  $t \geq t_1$ , characterized by increasing failure rate primarily attributable to wear-out (the human population analog is “old-age mortality”).



**FIGURE 21.6:** Typical failure rate (hazard function) curve showing the classic three distinct characteristic periods in the lifetime distributions of a population of items

In light of such characteristics, manufacturers often improve product reliability by (i) putting their batch of manufactured products through an initial “burn in” period of pre-release operation until time  $t_0$  to weed out the inferior items, and (ii) where possible, by replacing (or at least recommending replacement) at  $t_1$  to avoid failure due to wear-out. For example, this is the rationale behind the 90,000-mile timing belt replacement recommendation for some automobiles.

### 21.3.2 Probability Models for Distribution of Failure Times

One of the primary utilities of the expression in Eq (21.37) is that given any hazard function (failure rate), the corresponding distribution of failure times can be obtained directly. For example, for random chance failures, with constant failure rate,  $\eta$ , this equation immediately yields:

$$f(t) = \eta e^{-\eta t} \quad (21.38)$$

recognizable as the exponential pdf. We may now recall the discussion in Chapter 9 where the waiting time to the first occurrence of a Poisson event was identified as an exponential random variable. If component failure is a Poisson event, then Eq (21.38) is consistent with that earlier discussion.

Thus, for constant failure rate lifetime models, the failure-time distribution,  $f(t)$ , is exponential, with parameter  $\eta$  as the failure rate. The mean-time-to-failure is then  $1/\eta$ . If the component is replaced upon failure with an identical one (with the same constant failure rate,  $\eta$ ), then  $1/\eta$  is known as the mean-time-between-failures (MTBF).

For components or systems with the exponential failure-time distribution

in Eq (21.38), the reliability function is:

$$R(t) = 1 - F(t) = e^{-\eta t}. \quad (21.39)$$

This reliability function is valid in the “normal” period of the product lifetime, the middle section of the failure rate curve.

During the “initial” and “final” periods of product lifetimes, the failure rates are not constant, decreasing in one and increasing in the other. A more appropriate failure-rate function is

$$h(t) = \zeta \eta (\eta t)^{\zeta-1}; t > 0 \quad (21.40)$$

a very general failure rate function: when  $\zeta < 1$  it represents a failure rate that decreases with time, the so-called decreasing failure rate (DFR) model;  $\zeta > 1$  represents an increasing failure rate (IFR) model; and when  $\zeta = 1$ , the failure rate is constant at  $\eta$ . This expression therefore covers all the three periods of Fig 21.6.

The corresponding pdf,  $f(t)$ , for this general hazard function is obtained from Eq (21.37) as

$$f(t) = \eta \zeta (\eta t)^{\zeta-1} e^{-(\eta t)^\zeta} \quad (21.41)$$

recognizable as the Weibull pdf. The reliability function is obtained from Eq (21.33) as:

$$R(t) = e^{-(\eta t)^\zeta} \quad (21.42)$$

so that the cdf is therefore:

$$F(t) = 1 - e^{-(\eta t)^\zeta} \quad (21.43)$$

## 21.4 The Exponential Reliability Model

### 21.4.1 Component Characteristics

If the failure rates of component  $C_i$  of a system can be considered constant, with value  $\eta_i$ , then

1.  $f_i(t)$ , the failure-time distribution, is exponential, from which,
2.  $R_i(t)$ , the component reliability function (as a function of time-in-service,  $t$ ) is obtained as,

$$R_i(t) = e^{-\eta_i t} \quad (21.44)$$

From here, we can compute  $R_s(t)$ , the reliability of a system consisting of  $n$  such components, once we are given the system configuration; and from  $R_s(t)$ , we can then compute the system's MTBF. Let us illustrate first what Eq (21.44) means with the following example.

**Example 21.6: TIME-DEPENDENCY OF COMPONENT RELIABILITY**

Consider a component with failure rate 0.02 per thousand hours; i.e.,  $\eta_i = 0.02/1000$ . Obtain  $R_i(1000)$  and  $R_i(5000)$ , the probabilities that the component will be in service respectively for at least 1000 hours, and for at least 5000 hours. Also obtain the MTBF.

**Solution:**

The probability that this component will be in service for *at least* 1000 hours is obtained as

$$R_i(1000) = e^{[-(0.02/1000) \times 1000]} = e^{-0.02} = 0.98 \quad (21.45)$$

The probability that the same item will remain in service for at least 5000 hours is

$$R_i(5000) = e^{[-(0.02/1000) \times 5000]} = e^{-0.1} = 0.905 \quad (21.46)$$

Thus, if the time-in-service is changed, the reliability will also change; and for components with exponential failure-time distributions, the time-dependency is represented by Eq (21.44). The longer the required time-in-service for these components the lower the reliability. In the limit as  $t \rightarrow \infty$ , the probability that such components remain in service goes to zero: (nothing lasts forever!)

For this component, the mean-time-between-failure is

$$MTBF_i = \frac{1}{\eta} = \frac{1000}{0.02} = 5 \times 10^4 \text{ hrs} \quad (21.47)$$

which is constant.

#### 21.4.2 Series Configuration

For a system consisting of  $n$  components in series, each with reliability  $R_i(t)$  as given above in Eq (21.44), the resulting system reliability is

$$\begin{aligned} R_s(t) &= \prod_{i=1}^n e^{-\eta_i t} = e^{-(\sum_{i=1}^n \eta_i)t} \\ &= e^{-\eta_s t} \end{aligned} \quad (21.48)$$

where

$$\eta_s = \sum_{i=1}^n \eta_i \quad (21.49)$$

Thus, the failure-time distribution of a series configuration of  $n$  components, each having an exponential failure-time distribution, is itself an exponential distribution.

The system's MTBF is obtained as follows: for a component with reliability given in Eq (21.44), the component MTBF, say  $\mu_i$ , is:

$$\mu_i = \frac{1}{\eta_i} \quad (21.50)$$

As such, for the system, with  $R_s$  as given in Eq (21.48), the MTBF,  $\mu_s$ , is given by:

$$\mu_s = \frac{1}{\eta_s} \quad (21.51)$$

with  $\eta_s$  as defined in Eq (21.49). Thus,

$$\mu_s = \frac{1}{\eta_1 + \eta_2 + \cdots + \eta_n} = \frac{1}{\frac{1}{\mu_1} + \frac{1}{\mu_2} + \cdots + \frac{1}{\mu_n}} \quad (21.52)$$

so that:

$$\frac{1}{\mu_s} = \frac{1}{\mu_1} + \frac{1}{\mu_2} + \cdots + \frac{1}{\mu_n} \quad (21.53)$$

i.e., the MTBF of a series configuration of  $n$  components, each with individual MTBF of  $\mu_i$ , is the harmonic mean of the component MTBF's.

In the special case where all the components are identical, (in which case,  $\eta_i = \eta$ ), and therefore  $\mu_i = \mu$ , then

$$\eta_s = n\eta \quad (21.54)$$

$$\mu_s = \frac{\mu}{n} \quad (21.55)$$

i.e., the system failure rate is  $n$  times the component failure rate, and the MTBF is  $1/n$  times that of the component.

### 21.4.3 Parallel Configuration

The reliability  $R_s$  of a system consisting of  $n$  components in parallel, each with reliability  $R_i(t)$  as given above in Eq (21.44), is given by:

$$R_s(t) = 1 - \prod_{i=1}^n (1 - e^{-\eta_i t}) \quad (21.56)$$

which is *not* the reliability function for an exponential failure-time distribution. In the special case where the component failure rates are identical, this expression simplifies to,

$$R_s(t) = 1 - (1 - e^{-\eta t})^n \quad (21.57)$$

In general, the expression for the MTBF is difficult to derive, but it can be shown that it is given by:

$$\mu_s = \mu \left( 1 + \frac{1}{2} + \cdots + \frac{1}{n} \right) \quad (21.58)$$

and the system failure rate by

$$\frac{1}{\eta_s} = \frac{1}{\eta} \left( 1 + \frac{1}{2} + \cdots + \frac{1}{n} \right) \quad (21.59)$$

Some important implications of these results for the parallel system configuration are as follows:

1. The MTBF for a system of  $n$  identical components in parallel is the indicated series sum to  $n$  terms multiplied by the individual component MTBF. Keep in mind that this assumes that each defective component is replaced when it fails (otherwise  $n$  cannot remain constant).
2. In going from a single component to two in parallel, the MTBF increases by a factor of 50% (from  $\mu$  to  $1.5\mu$ ), not 100%.
3. The law of “diminishing returns” is evident in Eq (21.59): as far as MTBF for a parallel system configuration is concerned, the incremental benefits accruing from adding one more component to the system, goes to zero as  $n \rightarrow \infty$ .

#### 21.4.4 $m$ -of- $n$ Parallel Systems

When the system status depends on  $m$  components all functioning, then it will take multiple sequential failures ( $n - m = k$  of such) for the entire system to fail. If each component is independent, and each has an exponential failure-time distribution with common failure rate  $\eta$ , then it can be shown quite straightforwardly (recall the discussions in Chapter 9) that the system failure-time distribution is the gamma distribution:

$$f_s(t) = \frac{\eta^k}{\Gamma(k)} e^{-\eta t} t^{k-1} \quad (21.60)$$

with  $k = n - m$ . This is the waiting time to the occurrence of the  $k^{th}$  Poisson event when these events are occurring at a mean rate  $\eta$ . (This can also be written explicitly in terms of  $m$  and  $n$  simply by replacing  $k$  with  $(n - m)$ .) Since this is the pdf of a gamma  $\gamma(k, 1/\eta)$  random variable, whose mean value is therefore  $k/\eta$ , we immediately obtain the MTBF for this system as:

$$MTBF_s = \mu_s = \frac{k}{\eta} = (n - m)\mu \quad (21.61)$$

## 21.5 The Weibull Reliability Model

As noted earlier, the exponential reliability model is only valid for constant failure rates. When failure rates are time dependent, the Weibull model is more appropriate. Unfortunately, the Weibull model, being quite a bit more complicated than the exponential model, does not lend itself as easily to the sort of closed form analysis presented above for the exponential counterpart.

The component reliability is given from Eq (21.42) by:

$$R_i(t) = e^{-(\eta_i t)^\zeta} \quad (21.62)$$

and when the failure rate exponent  $\zeta$  is assumed to be the same for  $n$  components connected in *series*, the resulting system reliability is:

$$\begin{aligned} R_s(t) &= \prod_{i=1}^n e^{-(\eta_i t)^\zeta} = e^{-(\sum_{i=1}^n \eta_i^\zeta) t^\zeta} \\ &= e^{(-\eta_s t)^\zeta} \end{aligned} \quad (21.63)$$

where:

$$\eta_s = \left( \sum_{i=1}^n \eta_i^\zeta \right)^{1/\zeta} \quad (21.64)$$

Thus, as with the exponential case, the failure-time distribution of a series configuration of  $n$  components, each having a Weibull failure-time distribution,  $W(\eta_i, \zeta)$  (i.e., with identical  $\zeta$ ), is itself another Weibull distribution,  $W(\eta_s, \zeta)$ . If  $\zeta$  is different for each component, the system reliability function is quite a bit more complicated.

From the characteristics of the Weibull random variable, the component MTBF is obtained as:

$$\mu_i = \frac{1}{\eta_i} \Gamma(1 + 1/\zeta) \quad (21.65)$$

As such, the system MTBF with  $R_s$  as in Eq (21.63) is:

$$\mu_s = \frac{1}{\eta_s} \Gamma(1 + 1/\zeta) = \frac{\Gamma(1 + 1/\zeta)}{\left( \sum_{i=1}^n \eta_i^\zeta \right)^{1/\zeta}} \quad (21.66)$$

again, provided that  $\zeta$  is common to all components, because only then is Eq (21.64) valid.

The system reliability function,  $R_s(t)$  for the *parallel* configuration is

$$R_s(t) = 1 - \prod_{i=1}^n \left[ 1 - e^{(-\eta_i t)^\zeta} \right] \quad (21.67)$$

and even with  $\eta_i = \eta$  and a uniform  $\zeta$ , the expression is quite complicated, and must be computed numerically.

## 21.6 Life Testing

The experimental procedure for determining component and system reliability and lifetimes parallels the procedure for statistical inference discussed in Part IV: it involves selecting an appropriate random sample of components and testing them under prescribed conditions. The relevant data are the times-to-failure observed for individual components of system. Such experiments are usually called *life tests* and the general procedure is known as *life testing*. There are several different types of life tests, a few of the most common of which are listed below:

1. *Replacement tests*: where each failing component is replaced by a new one immediately upon failure;
2. *Nonreplacement tests*: where a failing component is *not* replaced;
3. *Truncated tests*: where, because the mean lifetime is so long that testing to failure is impractical, uneconomical, or both, the test is stopped (truncated) after (i) a fixed pre-specified time, or (ii) the first  $r < n$  failures;
4. *Accelerated tests*: where high-reliability components are tested under conditions far more severe than normal, in order to accelerate component failure and thereby reduce test time and the total number of components to be tested. The true natural reliability is extracted from such accelerated tests via standard analysis tools calibrated for the implied time-compression.

Once more, we caution the reader that the ensuing abbreviated discussion is meant to be merely illustrative, nowhere near the fuller, more comprehensive discussion of the fundamental principles and results that are available in such book-length treatments as that in Nelson, 2003<sup>1</sup>.

### 21.6.1 The Exponential Model

As shown earlier in Section 21.4, the exponential model is the most appropriate lifetime model during the useful life period. The main feature of the life tests for this model is that  $n$  components are life-tested independently and testing is discontinued after  $r \leq n$  have failed. The experimental result is the set of observed failure times:  $t_1 \leq t_2 \leq t_3 \cdots \leq t_r$ , where  $t_i$  is the failure time of the  $i^{\text{th}}$  component to fail.

Statistical inference in this case involves the usual problems: *estimation* of the key population parameter,  $\mu = 1/\eta$  of the exponential failure-time

<sup>1</sup>Nelson, W. B. (2003). *Applied Life Data Analysis*, Wiley, NY.

distribution, the mean component lifetime; and *hypothesis testing* about the parameter estimate, but with a twist.

### Estimation

It can be shown that an unbiased estimate for  $\mu$  is given by:

$$\hat{\mu} = \frac{\tau_r}{r} \quad (21.68)$$

where  $\tau_r$  is the “accumulated life until the  $r^{th}$  failure” given by:

$$\tau_r = \sum_{i=1}^r t_i + (n - r)t_r \quad (21.69)$$

for non-replacement tests. The first term is the total lifetime of the  $r$  failed components; the second term is the lower bound on the remaining lifetime of the surviving  $(n - r)$  components. Note that for these non-replacement tests, if  $r = n$ , then  $\hat{\mu}$  is exactly equal to the mean of the observed failure times. For replacement tests,

$$\tau_r = nt_r \quad (21.70)$$

From here, the failure rate is estimated as

$$\hat{\eta} = 1/\hat{\mu} \quad (21.71)$$

and the reliability as

$$\hat{R}(t) = e^{-t/\hat{\mu}} \quad (21.72)$$

It can be shown that these estimates are biased, but the bias diminishes as the sample size  $n$  increases.

### Precision of Estimates

The statistic

$$W_r = \frac{2\mathcal{T}_r}{\mu} \quad (21.73)$$

where  $\mathcal{T}_r$  is the random variable whose specific value,  $\tau_r$ , is given in Eq (21.69) or (21.70), possesses a  $\chi^2(2r)$  distribution. As a result, the usual two-sided  $(1 - \alpha) \times 100\%$  confidence interval for  $\mu$  may be obtained from the fact that

$$P \left[ \chi^2_{1-\alpha/2}(2r) < W_r < \chi^2_{\alpha/2}(2r) \right] = 1 - \alpha \quad (21.74)$$

following precisely the same arguments as in Chapter 14. The result is that:

$$\frac{2\mathcal{T}_r}{\chi^2_{\alpha/2}(2r)} < \mu < \frac{2\mathcal{T}_r}{\chi^2_{1-\alpha/2}(2r)} \quad (21.75)$$

represents the  $(1 - \alpha) \times 100\%$  confidence interval for  $\mu$ .

**TABLE 21.1:** Summary of  $H_0$   
rejection conditions for the test of  
hypothesis based on an exponential model  
of component failure-time

| Testing Against        | For general $\alpha$<br>Reject $H_0$ if:                                      |
|------------------------|---|
| $H_a : \mu < \mu_0$    | $\omega_r < \chi_{1-\alpha}^2(2r)$  |
| $H_a : \mu > \mu_0$    | $\omega_r > \chi_\alpha^2(2r)$  |
| $H_a : \mu \neq \mu_0$ | $\omega_r < \chi_{1-\alpha/2}^2(2r)$ or<br>$\omega_r > \chi_{\alpha/2}^2(2r)$ |

### Hypothesis Testing

To test the null hypothesis

$$H_0 : \mu = \mu_0$$

against the usual triplet of alternatives:

$$\begin{aligned} H_a : & \quad \mu > \mu_0 \\ H_a : & \quad \mu < \mu_0 \\ H_a : & \quad \mu \neq \mu_0 \end{aligned}$$

again, we follow the principles discussed in Chapter 15. We use the test statistic  $W_r$  defined in Eq (21.73) and its sampling distribution, the  $\chi^2(2r)$  distribution, to obtain the usual rejection criteria, shown for this specific case in Table 21.1, where  $\omega_r$  is the specific value of the statistic obtained from experimental data, i.e.,

$$\omega_r = \frac{2\tau_r}{\mu_0} \tag{21.76}$$

Even though all these closed-form results are available, none of these statistical inference exercises are conducted by hand any longer. As with the examples discussed in Chapters 14 and 15, computer programs are routinely used for such data analysis. Nevertheless, we use the next example to illustrate some of the mechanics behind the computations.

#### Example 21.7: LIFE TESTS FOR ENERGY-SAVING LIGHT BULBS

To characterize the lifetime of a new brand of energy-saving light bulbs, a sample of 10 were tested in a specially designed facility where they could be left on continuously and monitored electronically to record the precise number of hours until burn out. The experimental design calls for halting the test immediately after 8 of the 10 light bulbs have burned

out. The result, in thousands of hours, arranged in increasing order is as follows:

$$(1.599, 3.380, 5.068, 8.478, 8.759, 9.256, 11.475, 14.382)$$

i.e., the first light bulb to burn out did so after 1,599 hours, the next after 3,380 hours, and the 8<sup>th</sup> and final one after 14,382 hours. Obtain an estimate of the mean lifetime and test the hypothesis that it is 12,000 hours against the alternative that it is not.

**Solution:**

For this problem,

$$\tau_r = 62.397 + (10 - 8) \times 14.382 = 91.161 \quad (21.77)$$

so that the estimate for  $\mu$  is:

$$\hat{\mu} = 11.40 \quad (21.78)$$

The test statistic,  $\omega_r$  is:

$$\omega_r = 182.322/12 = 15.194 \quad (21.79)$$

And from the chi-square distribution, we obtain  $\chi^2_{0.025}(16) = 6.91$  and  $\chi^2_{0.975}(16) = 28.8$ . And now, since 15.194 does not lie in the rejection region, we find no evidence to reject the null hypothesis. We therefore conclude that it seems reasonable to assume that the true mean lifetime of this new brand of light bulb is 12,000 hours as specified.

### 21.6.2 The Weibull Model

When the failure rate function is either decreasing or increasing, as is the case in the initial and final periods of component lifetimes, the Weibull model is more appropriate. The failure-time distribution, reliability function, and failure rate (or hazard function) for this model were given earlier. From the discussions in Chapter 8, we know that the mean of the Weibull pdf, which in this case will correspond to the mean failure time, is given by:

$$\mu = E(T) = \frac{1}{\eta} \Gamma(1 + 1/\zeta) \quad (21.80)$$

Life testing is aimed at acquiring data from which the population parameters,  $\zeta$  and  $\eta$  will be estimated. Unfortunately, unlike with the exponential model, estimating these Weibull parameters can be tedious and difficult, requiring either numerical methods or old-fashioned graphical techniques that are based on many simplifying approximations. Even more so than with the relatively simpler exponential model case, computer software must be employed for carrying out parameter estimation, and hypothesis tests for the Weibull model.

Additional details lie outside the intended scope of this chapter but are available in the book by Nelson (2003), which is highly recommended to the interested reader.

## 21.7 Summary and Conclusions

The exposure to the topic of reliability and life testing provided in this chapter was designed to serve two purposes. First is the general purpose of Part V—to showcase, no matter how briefly, some substantial subject matters that are based entirely on applications of probability and statistics. Second is the specific purpose of illustrating how the reliability and the lifetimes of components and systems are characterized and analyzed. The scope of coverage was deliberately limited, but still with the objective of providing enough material such that the reader can develop a sense of what these studies entail. We presented reliability, for a component or a system, as a probability—the probability that the component or system functions as desired, for at least a specified period of time. The techniques discussed for determining system reliability given component reliabilities and system configuration produced some interesting results, two of which are summarized below:

- *Product Law of Reliabilities:* For systems consisting of  $n$  components connected in series, each with reliability,  $R_i$ , the system reliability,  $R_s$ , is a product of the component reliabilities; i.e.,  $R_s = \prod_{i=1}^n R_i$ . Since  $0 < R_i < 1$ , the reliability of a system of series components therefore diminishes as the number of components increases.
- *Product Law of Unreliabilities:* When the  $n$  components of a system are arranged in parallel, the system *unreliability*,  $(1 - R_s)$ , is a product of the component unreliabilities; i.e.,  $1 - R_s = \prod_{i=1}^n (1 - R_i)$ . Thus, the reliability of a system of parallel components improves as the number of components increases; the additional components simply act as redundant backups.

Computing the reliabilities of more complex systems requires reducing such systems to a collection of simple modules and, in the case of systems with cross-links, using a keystone and invoking Bayes' theorem of total probability.

As far as specific models of failure times are concerned, we focused only on the exponential and Weibull models, the two most widely used in practice. In discussing failure time distributions and their characteristics, we were able to revisit some of the special lifetime distributions presented earlier in Chapter 4 (especially the survival function and the hazard function) here in their more natural habitats.

While reliability analysis depends entirely on probability, not surprisingly, life testing, the experimental determination of component and system reliability characteristics, relies on statistical inference: estimation and hypothesis testing. How these ideas are used in practice is illustrated further with the end-of-chapter exercises and problems.

**REVIEW QUESTIONS**

- 1.** What is the definition of the reliability of a component or a system?
- 2.** What are the two factors that determine the overall reliability of a system consisting of several components?
- 3.** What is the defining problem of system reliability?
- 4.** In terms of system configuration, what is a “simple” system as opposed to a “complex” system?
- 5.** What is the product law of reliabilities, and to which system configuration does it apply?
- 6.** Why is system reliability for a series configuration independent of the order in which the components are arranged?
- 7.** What is the product law of unreliabilities, and to which system configuration does it apply?
- 8.** As  $n$ , the number of components in a series configuration increases, what happens to  $R_s$ , system reliability? Does it increase or decrease?
- 9.** As  $n$ , the number of components in a parallel configuration increases, what happens to  $R_s$ , system reliability? Does it increase or decrease?
- 10.** What is a module?
- 11.** What is the analysis technique for determining the reliability of series-parallel systems?
- 12.** What is a  $k$ -of- $n$  parallel system?
- 13.** Why is it more complicated than usual to determine the reliability of systems with cross-links?
- 14.** What special component designation is needed in analyzing the reliability of systems with cross-links?
- 15.** What is the survival function,  $S(t)$ , and how is it related to the cumulative distribution function,  $F(t)$ ?
- 16.** In lifetime studies, the cumulative distribution function,  $F(t)$ , is the same as what system characteristic?
- 17.** What is the hazard function,  $h(t)$ , and how is it related to the standard pdf,  $f(t)$ ?

- 18.** Why is the failure rate (hazard function) curve known as the bathtub curve?
- 19.** The “rate” in the failure rate is not with respect to time; it is with respect to what?
- 20.** What are the three distinct parts of the hazard function (failure rate) curve?
- 21.** What is the distribution of failure times for random chance failure, with constant failure rates,  $\eta$ ?
- 22.** What is the reliability function for components or systems with exponential failure-time distributions?
- 23.** What is the definition of mean-time-between-failures (MTBF)?
- 24.** What is a decreasing failure rate (DFR) as opposed to an increasing failure rate (IFR) model?
- 25.** What is the reliability function for components or systems with the Weibull failure-time distribution?
- 26.** What is the failure time distribution for a series configuration of  $n$  components each with an exponential failure-time distribution?
- 27.** What is the relationship between the MTBF of a series configuration of  $n$  components each with exponential failure-time distributions, and the MTBFs of the components?
- 28.** In what way is the law of diminishing returns manifested in the MTBF for a parallel configuration of  $n$  identical systems with exponential reliability?
- 29.** What is the MTBF for an  $m$ -of- $n$  parallel system with exponential component reliabilities?
- 30.** What is the failure-time distribution for a series configuration of  $n$  components each with a Weibull failure-time distribution?
- 31.** What is life testing?
- 32.** What is a replacement test, a non-replacement test, a truncated test, or an accelerated test?
- 33.** In life testing, what is the “accumulated life until the  $r^{th}$  failure”?
- 34.** What test statistic is used in life testing with the exponential model? What is its sampling distribution?

## EXERCISES AND APPLICATION PROBLEMS

**21.1** The facility for storing blood in a blood bank located in a remote hospital consists of a primary refrigerator  $RF_1$ , a backup,  $RF_2$ , and a set of gasoline engine generators,  $G_1$  the primary generator,  $G_2$  the first backup and  $G_3$  yet another backup. The primary system is  $RF_1$  connected to  $G_1$  with  $G_2$  as backup; the entire primary system is backed up by  $RF_2$  connected to  $G_3$ . as shown in Fig 21.7.

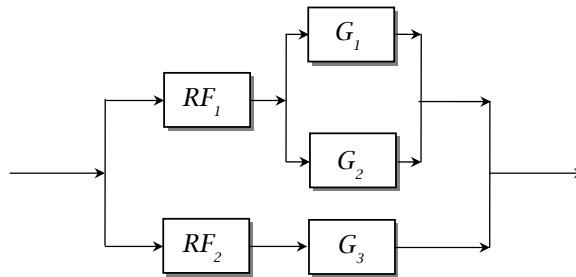


FIGURE 21.7: Blood storage system

- (i) If the refrigerators are identical, with reliabilities,  $R = 0.95$ , and the generators have the following reliabilities:  $R_{G_1} = 0.85 = R_{G_2}$ , and  $R_{G_3} = 0.75$ , compute the overall system reliability.
- (ii) It has been argued that the backup generator  $G_3$ , also needs a back-up (just like  $G_2$  was a backup for  $G_1$ ) so that the full backup system will replicate the full primary system exactly. If a generator  $G_4$  with the same reliability as that of  $G_3$  is added as prescribed, what is the resulting percentage increase in overall system reliability?

**21.2** A single distributed control system (DCS) used to carry out all the control functions in a manufacturing facility has a reliability of 0.99. DCS manufacturers usually recommend purchasing not just a single system, but a complete system with built-in back-up modules that function in standby mode in parallel with the primary module. (i) If it is desired to have a complete system with a reliability of 0.999, at least how many modules in parallel will be required to achieve this objective? (ii) If the reliability is increased to 0.9999 how many modules will be required?

**21.3** The following reliability block diagram is for a heat exchange system employed in a nuclear power generating plant. Three heat exchangers  $HX_1$ ,  $HX_2$  and  $HX_3$  are each equipped with control valves and ancillary control systems for maintaining temperature control. The reliability,  $R = 0.9$ , associated with each heat exchanger is due to the possible failure of the control valves/control system assembly. For simplicity, these components are not shown explicitly in the diagram. The three heat exchangers are supplied with cooling water by three pumps,  $P_1$ ,  $P_2$  and  $P_3$ , each with reliability 0.95. The pump characteristics are such that any one of them is sufficient to service all three heat exchangers; however, the power plant requires 2 of the three heat exchangers to function. The water supply may be assumed to be always reliable.

- (i) Determine the system reliability under these conditions.

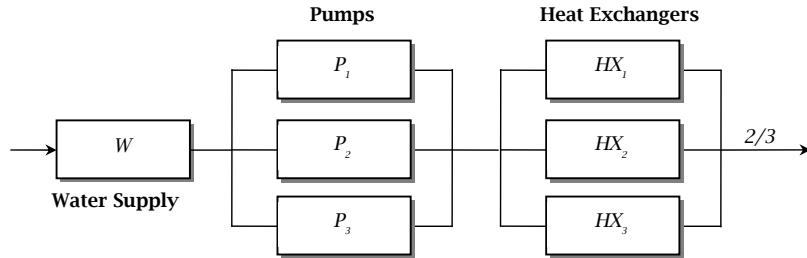


FIGURE 21.8: Nuclear power plant heat exchanger system

- (ii) If the power plant were redesigned such that only one of the heat exchangers is required, by how much will the system reliability increase?

**21.4**  $R_s$ , the reliability of the system shown below, was obtained in the text using component  $C_3$  as the “keystone.”

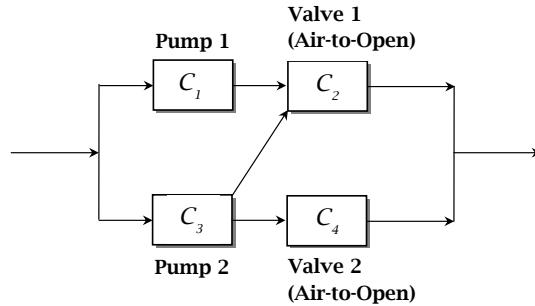


FIGURE 21.9: Fluid flow system with a cross link (from Fig 21.5)

- (i) Choose  $C_2$  as the keystone and obtain  $R_s$  again. Compare your result with Eqn (21.30) in the text.  
(ii) Now choose  $C_1$  as the keystone and repeat (i). Compared with the derivation required in (i), which keystone choice led to a more straightforward analysis?  
(iii) Given specific component reliabilities for the system as:  $R_1 = 0.93$ ;  $R_2 = 0.99$ ;  $R_3 = 0.93$ ;  $R_4 = 0.99$ , where  $R_i$  represents the reliability of component  $C_i$ , compare the reliability of the system with and without the cross-link and comment on how the presence of the cross-link affects this specific system's reliability.

**21.5** An old-fashioned fire alarm system consists of a detector  $D$  and an electrically operated bell,  $B$ . The system works as follows: if a fire is detected, a circuit is completed and the electrical signal reaching the bell will cause it to ring. The reliability of the detector is 0.9 and that of the bell is 0.995.

- (i) What is the reliability of the complete fire alarm system?  
(ii) If another identical detector and bell combination is installed in standby, by how much will the reliability of the new augmented fire alarm system improve?  
(iii) It has been recommended, as a cost-saving measure, to purchase for the back-

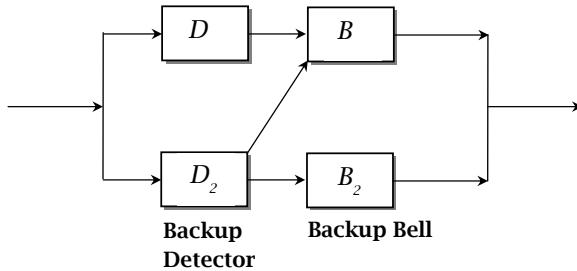


FIGURE 21.10: Fire alarm system with back up

up system proposed in (ii), a detector  $D_2$ , that is identical to the primary detector, but a less reliable bell  $B_2$  ( $R = 0.85$ ) and wire the overall system such that  $D_2$  is connected to the primary bell  $B$  via a cross-link, as shown in the diagram below, Fig 21.10. Determine the reliability of the new system with this configuration.

**21.6** Refer to Problem 21.5, part (iii). Determine the system reliability if the system wiring is adjusted so that in addition, a second cross-link connects the primary detector  $D$  to the less reliable back-up bell,  $B_2$ . Does the system reliability increase or decrease?

**21.7** The condenser system used to condense volatile organic compounds (VOCs) out of a vapor stream in a petrochemical facility consists of a temperature sensor,  $S$ , an electronic controller,  $C$ , and a heat exchanger,  $HX$ . So long as the temperature of the vapor stream through the condenser is maintained by the control system below the boiling point of the lowest boiling VOC, there will be no release of VOCs to the atmosphere. The reliability of the sensor is  $R_S$ , that of the controller is  $R_C$ , and the entire heat exchanger ensemble has a reliability of  $R_{HX}$ . The following configuration is used for the particular system in question, a full scale back up with cross-links.

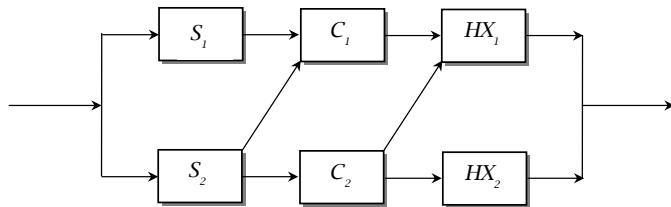
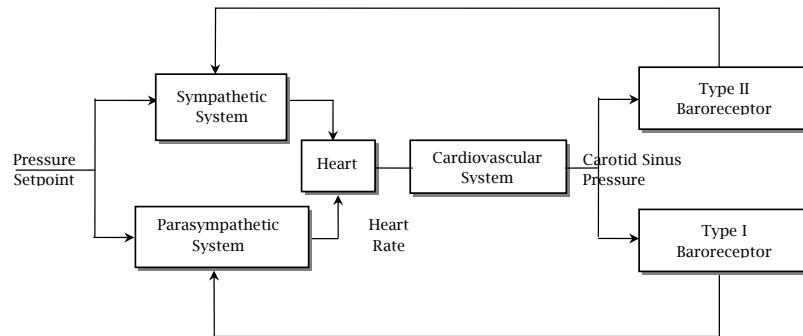


FIGURE 21.11: Condenser system for VOCs

- Determine the system reliability, assuming that similar components have identical reliabilities.
- Given the following component reliabilities,  $R_S = 0.85$ ;  $R_C = 0.9$ ;  $R_{HX} = 0.95$ , determine the overall system reliabilities.

**21.8** Pottmann. *et al.*, (1996)<sup>2</sup> presented the following simplified block diagram for the mammalian blood pressure control system. The baroreceptors are themselves systems of pressure sensors, and the sympathetic and parasympathetic systems are separate control systems that are part of the nervous systems. These subsystems are not entirely all mutually independent; for the purposes of this problem, however, they can be considered as such. Consider an experimental rat for which the indicated



**FIGURE 21.12:** Simplified representation of the control structure in the baroreceptor reflex

baroreceptor reflex components have the following reliabilities:

- Sensors: Baroreceptor Type I,  $R_{B1} = 0.91$ ; Type II,  $R_{B2} = 0.85$ ;
- Controllers: Sympathetic system,  $R_S = 0.95$ ; Parasympathetic system,  $R_P = 0.92$ ;
- Actuators: Heart,  $R_H = 0.99$ ; Cardiovascular system,  $R_C = 0.95$

Rearrange the control system block diagram into a reliability block diagram by starting with the two systems representing the sensors at the left end (and the carotid sinus pressure signal as the input), then connect the sensors in parallel to the appropriate controllers, (the pressure setpoints to the controllers may be omitted for the reliability analysis), ultimately end with the carotid sinus pressure signal from the cardiovascular system.

- Determine the system reliability as configured.
- If the system operated only via the Type I receptors-Parasympathetic system connection, determine the system reliability under these conditions.
- Repeat (ii) under the conditions that the system operated only via the alternative Type II receptors-Sympathetic system connections. Which of the isolated systems is more reliable? By how much is the overall system reliability improved as a result of the parallel arrangement?

**21.9** In designing a safety system for a high pressure ethylene-copolymers manufacturing process, the design engineers had to take the following factors into consideration: First, no safety system, no matter how sophisticated, can be perfect in

<sup>2</sup>Pottmann, M., M. A. Henson, B. A. Ogunnaike, and J. S. Schwaber, (1996). "A parallel control strategy abstracted from the baroreceptor reflex," *Chemical Engineering Science*, 51 (6), 931-945.

preventing accidents; the higher the safety system reliability, the lower the risk, but it will never be zero. Second, attaining high system reliability is not cheap, whether it is realized with individual components with high reliability, or by multiple redundancies. Last, but not least, even though high reliability is expensive, the repercussions of a single safety catastrophe with this manufacturing process is enormous in financial terms, besides the lingering effects of bad publicity that can take decades to overcome, if ever.

Engineers designing a safety system for a specific plant were therefore faced with a difficult optimization problem: balancing the high cost of a near-perfect system against the enormous financial repercussions of a single catastrophic safety event. But an optimum solution can be obtained as follows.

Let  $C_0$  be the cost of a reference system with mediocre reliability of 0.5 (i.e., a system with a 50% probability of failure). For the particular process in question, the total cost of installing a system with reliability  $R_s$  is given by:

$$\phi_R = \frac{C_0 R_s}{1 - R_s} \quad (21.81)$$

Note that in the limit as  $R_s \rightarrow 1$ ,  $\phi_R \rightarrow \infty$ . There is a point of diminishing returns, however, above which the incremental cost does not result in commensurate increase in reliability; i.e., eking out some additional increase in reliability beyond a certain value is achieved at a disproportionately high cost.

Now let the cost of system failure be  $C_F$ , which, by definition, is a substantial sum of money. For a system with reliability  $R_s$ , the expected cost of failure is,

$$\phi_F = C_F(1 - R_s) \quad (21.82)$$

where, of course,  $(1 - R_s)$  is the probability of system failure. Note that, as expected, this is a monotonically (specifically, linearly) decreasing function of system reliability.

We may now observe that the ideal system will be one with a reliability that minimizes the total expected costs, achieving a high enough reliability to reduce the risk of failure, but not so much that the cost of reliability is prohibitive.

(i) Determine such a reliability,  $R^*$ , by minimizing the objective

$$\phi = \phi_F + \phi_R = C_F(1 - R_s) + \frac{C_0 R_s}{1 - R_s} \quad (21.83)$$

the total expected costs; show that the desired optimum reliability is given by:

$$R^* = 1 - \sqrt{\frac{C_0}{C_F}} = 1 - \sqrt{\rho} \quad (21.84)$$

where  $\rho$  is the ratio of the base reference cost of reliability to the cost of failure. Discuss why this result makes sense by examining the prescribed  $R^*$  as a function of the ratio  $\rho$  of the two costs,  $C_0$  and  $C_F$ .

- (ii) For a particular system where  $C_0 = \$20,000$  and for which  $C_F = \$500$  million, determine  $R^*$ , and the cost of the safety system whose reliability is  $R^*$ .
- (iii) If the cost of the catastrophe were to double, determine the new value for  $R^*$ , and the corresponding cost of the recommended safety system.
- (iv) If a single composite system with reliability  $R^*$  determined in (ii) above is unavailable, but only a system with reliability 0.85, how many of the available systems will be required to achieve the desired reliability? How should these be configured?

**21.10** For certain electronic components, survival beyond an initial period from  $t = 0$  to  $t = \tau$  is most crucial because thereafter, the failure rate becomes virtually negligible. For such cases, the hazard function (i.e., the failure rate) may be approximated as follows:

$$h(t) = \begin{cases} \eta(1 - t/\tau) & 0 < t < \tau \\ 0 & \text{otherwise} \end{cases} \quad (21.85)$$

Obtain an expression for  $f(t)$ , the failure time distribution, and the corresponding cumulative distribution,  $F(t)$ . From these results show that for such electronic components, the reliability function is given by:

$$R(t) = e^{-\eta t(1-t/2\tau)}$$

during the initial period,  $0 < t < \tau$ , and that thereafter (for  $t > \tau$ ), it is:

$$R(t) = e^{-\eta\tau/2}$$

**21.11** The time to failure,  $T$ , of an electronic component is known to be an exponentially distributed random variable with pdf

$$f(t) = \begin{cases} \eta e^{-\eta t}; & 0 < t < \infty \\ 0; & \text{elsewhere} \end{cases}$$

where the “failure rate,”  $\eta = 0.075$  per 100 hours of operation.

(i) If the component “reliability function”  $R_i(t)$  is defined as

$$R_i(t) = P(T > t) \quad (21.86)$$

the probability that the component functions at least up until time  $t$ , obtain an explicit expression for  $R_i(t)$  for this electronic component.

(ii) A system consisting of two of such components in *parallel* functions if at least one of them functions; again assuming that both components are identical, find the system reliability  $R_p(t)$  and compute  $R_p(1000)$ , the probability that the system survives at least 1,000 hours of operation.

**21.12** The failure time (in hours) for 15 electronic components is given below:

|       |       |       |       |       |       |       |      |
|-------|-------|-------|-------|-------|-------|-------|------|
| 337.0 | 290.5 | 219.7 | 739.2 | 900.4 | 36.7  | 348.6 | 44.7 |
| 408.9 | 183.4 | 174.2 | 330.8 | 102.2 | 731.4 | 73.5  |      |

- (i) First confirm that the data is reasonably exponentially distributed and then obtain an estimate of mean life time.
- (ii) The company that manufactures the electronic components claims that the mean life time is 400 hours. Test this hypothesis against the alternative that the mean life-time is lower. What is the conclusion of this test?
- (iii) Using the estimated mean life time to determine the exponential population mean failure rate,  $\eta$ , compute the probability that a system consisting of two of these components in parallel functions beyond 400 hours.

**21.13** Refer to Problem 12.12. This time, consider that the life test was stopped, by design, after 500 hours. Repeat the entire problem and compare the results. How close to the full data results are the results from the truncated test?



# Chapter 22

## Quality Assurance and Control

|        |   |     |
|--------|---|-----|
| 22.1   | Introduction .....  | 934 |
| 22.2   | Acceptance Sampling .....                                       | 935 |
| 22.2.1 | Basic Principles .....  | 936 |
|        | Basic Characteristics of Sampling Plans .....                   | 936 |
| 22.2.2 | Determining a Sampling Plan .....                               | 937 |
|        | The Operating Characteristic (OC) Curve .....                   | 938 |
|        | Approximation Techniques .....                                  | 940 |
|        | Characteristics of the OC Curve .....                           | 942 |
|        | Other Considerations .....                                      | 943 |
| 22.3   | Process and Quality Control .....                               | 943 |
| 22.3.1 | Underlying Philosophy .....                                     | 944 |
| 22.3.2 | Statistical Process Control .....                               | 944 |
| 22.3.3 | Basic Control Charts .....                                      | 946 |
|        | The Shewhart Xbar Chart .....                                   | 947 |
|        | The S-Chart .....   | 948 |
|        | Variations to the Xbar-S Chart: Xbar-R, and I & MR Charts ..... | 950 |
|        | The P-Chart .....   | 954 |
|        | The C-Chart .....   | 957 |
| 22.3.4 | Enhancements .....  | 958 |
|        | Motivation .....  | 958 |
|        | Western Electric Rules .....                                    | 959 |
|        | CUSUM Charts .....  | 960 |
|        | EWMA Charts .....   | 961 |
| 22.4   | Chemical Process Control .....                                  | 964 |
| 22.4.1 | Preliminary Considerations .....                                | 964 |
| 22.4.2 | Statistical Process Control (SPC) Perspective .....             | 965 |
| 22.4.3 | Engineering/Automatic Process Control (APC) Perspective .....   | 965 |
| 22.4.4 | SPC or APC .....  | 967 |
|        | When SPC is More Appropriate .....                              | 968 |
|        | When APC is More Appropriate .....                              | 968 |
| 22.5   | Process and Parameter Design .....                              | 969 |
| 22.5.1 | Basic Principles .....  | 969 |
| 22.5.2 | A Theoretical Rationale .....                                   | 970 |
| 22.6   | Summary and Conclusions .....                                   | 971 |
|        | REVIEW QUESTIONS .....  | 972 |
|        | PROJECT ASSIGNMENTS .....                                       | 975 |
|        | 1. Tracking the Dow .....                                       | 975 |
|        | 2. Diabetes and Process Control .....                           | 976 |
|        | 3. C-Chart for Sports Team .....                                | 976 |

*Find out the cause of this effect  
or rather say, the cause of this defect;  
for this effect defective comes by cause.*

William Shakespeare (1564–1616); *Hamlet, II, ii, 101*

Mass production, a uniquely 20<sup>th</sup> century invention, unquestionably transformed industrial productivity by making possible the manufacture of large quantities of products in a relatively short period of time. But making a product faster and making lots of it does not necessarily mean much if the product is not made *well*. If anything, making the product well every time and all the time, became a more challenging endeavor with the advent of mass production. It is only natural, therefore, that assuring the quality of mass produced goods has since become an integral part of any serious manufacturing enterprise. At first, *acceptance sampling* was introduced by customers to protect themselves from inadvertently receiving products of inferior quality and only discovering these defective items afterwards. Before a manufactured lot is accepted by the consumer, the strategy calls for a sample to be tested first, with the results of the test serving as a rational basis for deciding whether to accept the entire lot or to reject it. Producers later incorporated acceptance sampling into their product release protocols to prevent sending out inferior quality products. However, such an “after-the-fact” strategy was soon recognized as inefficient and, in the long run, too expensive. The subsequent evolution of quality assurance through *process and quality control* (where the objective is to identify causes of poor quality and correct them during production) to the total quality management philosophy of zero defects (which requires, in addition, the design of processes and process operating parameters to minimize the effect of uncontrollable factors on product quality) was rapid and inevitable.

A complete and thorough treatment of quality assurance and control requires more space than a single chapter can afford. As such, our objective in this chapter is more modestly set at providing an overview of the key concepts underlying three primary modes of quality assurance. We discuss first acceptance sampling, from the consumer’s as well as the producer’s perspectives; we then discuss in some detail, process and quality control, where the focus is on the manufacturing process itself. This discussion covers the usual terrain of statistical process control charts, but adds a brief section on engineering/automatic process control, comparing and contrasting the two philosophies. The final overview of Taguchi methods is quite brief, providing only a flavor of the concepts and ideas.

---

## 22.1 Introduction

In modern manufacturing processes with mass production capabilities, the primary objective is to make products that meet customer requirements as economically as possible. Ideally, the customer sets a specific desired target value,  $y^*$ , for a particular measurable indicator of product characteristic,  $Y$ , that the manufacturer must match; for example, a “Toughness” of 140 J/m<sup>3</sup> required of a certain polymer resin; ball bearings of 2 mm outer diameter;

computer laptop batteries with a lifetime of 5 years, etc. Because of inherent and unavoidable variability, however, the customer typically specifies in addition to the target value,  $y^*$ , a “tolerance limit,” say  $\pm\tau$ , so that the product is then said to meet customer requirements when the specific measured product characteristic value,  $y$ , lies in the interval  $y^* \pm \tau$ , otherwise it *does not* meet the objective.

Unavoidable variability in raw materials, in process and environmental conditions, in human operators, etc., eventually manifest as variability in the final product characteristics. The primary issue of concern to manufacturers is therefore *how to ensure the final product quality given such unavoidable variabilities*. Products that do not meet customer requirements are usually rejected, with adverse financial consequences to the producer; in addition, the psychological handicap of being associated with “inferior quality” can often be difficult to overcome.

Quality assurance and control is a subject matter devoted to the techniques and tools employed in modern manufacturing processes to ensure that products are manufactured to specification in spite of unavoidable variabilities. And few applications of probability and statistics have penetrated and transformed an industry to the extent that quality assurance and control has the manufacturing industry. The problems associated with assuring the quality of “mass produced” products may be categorized as follows:

1. Given a large lot of manufactured items, how do we assure product quality *before* sending them out (the producer’s concern) or before accepting them (the consumer’s concern)? This is the “acceptance sampling” problem, by definition, a post-production problem.
2. Given a production process and operating procedure, how do we operate the process to ensure that the resulting manufactured product meets the desired specifications? This is the “process control” problem, by definition, a during-production problem.
3. Can the production process be designed and the operating procedure formulated such that the resulting manufacturing operation is robust to the sort of intrinsic variabilities that propagate and ultimately translate into unacceptable variability in product quality? This is the “process (or parameter) design” problem, a pre-production problem.

Traditional quality assurance focused on Problem 1 then evolved to include Problem 2; more recently, Problem 3 has received greater attention as significant results from successful applications become widely publicized. In the ideal case, if Problem 3 is solved in the pre-production stage, and Problem 2 is handled effectively during production, then there will be fewer, if any, post-production rejections, and Problem 1 becomes a non-issue. This is the prevalent total quality management view of quality assurance. The rest of the chapter is organized around each of these problems in the order presented above.

## 22.2 Acceptance Sampling

### 22.2.1 Basic Principles

Acceptance sampling, as the name implies, is a procedure for sampling a batch of products to determine, objectively, whether or not to accept the whole batch on the basis of the information gathered from the sample. It has traditionally been considered as a procedure used by the customer to ascertain the quality of a procured lot before accepting it from the manufacturer; however, it applies equally well to the producer, who checks manufactured lots before sending them out to customers.

The defining characteristic is that for a large lot of items, exhaustive inspection is not only infeasible, it is also impractical and quite often unaffordable. This is especially so when the quality assurance test is destructive. The classic example is the very application that led to the development of the technique in the first place: U.S. military testing of bullets during WW II. Clearly, *every* bullet cannot be tested (as there will be nothing left for soldiers to use!) but without testing any at all, there is no way to ascertain which lot will perform as desired and which will not. As a result, the decision on lot quality must be based on the results of tests conducted on samples from the lot, making this a problem perfectly tailor-made for the application of statistical inference theory. A typical statement of the acceptance sampling problem is:

Let  $N$  be the total number of items in a manufactured lot for which  $\theta$  is the true but unknown fraction of defective (or non-conforming) items; a sample of  $n$  items drawn from the lot and tested; and  $x$  is the actual number of defective (or non-conforming) items found in the sample (a realization of the random variable,  $X$ ). The lot is accepted if  $x \leq c$ , a predetermined critical value, the “acceptance number”; otherwise the lot is rejected.

For a given lot size,  $N$ , the pair,  $(n, c)$  determines a sampling plan. Acceptance sampling therefore involves determining a sampling plan, implementing it to determine  $x$  and deciding whether to accept or reject the lot on the basis of how  $x$  compares with  $c$ . How  $n$ , the sample size, and  $c$ , the acceptance number, are determined is at the heart of the theory of acceptance sampling.

Before discussing how a sampling plan is generated, here are some basic concepts and terminology used in acceptance sampling.

#### Basic Characteristics of Sampling Plans

##### 1. Product Characterization by Attributes and Variables

The acceptability of a tested item is either based on a discrete (binary or count) criterion or on a continuous one. For example, acceptance is based on

a discrete *binary* criterion for a batch of electronic chips evaluated on the basis of whether the sampled items function or do not function. For a batch of silicon wafers evaluated on the basis of the *number* of flaws each contains, acceptance is based on a discrete *count* criterion. These discrete quantities are known as "attributes" and present what is known as an "acceptance sampling by attribute" problem. On the other hand, a batch of polymer resins evaluated on the basis of continuous measurements such as "Toughness," in  $J/m^3$ , or Density (in  $kg/m^3$ ), presents an "acceptance sampling by variable" problem. The batch is accepted only if the values obtained for these product variables lie in a prescribed range.

## 2. Acceptance and rejection criteria

A base line acceptance requirement for the fraction (or percent) of defective items in a manufactured lot is known as the Acceptable Quality Level (AQL). This is the value  $\theta_0$  such that if the fraction of defectives found in the lot after inspection,  $x/n$ , is such that  $x/n \leq \theta_0$ , then the lot is *definitely* acceptable. The complementary quantity is the Rejectable Quality Level (RQL) (also known as the Lot Tolerance Percent Defective (LTPD)): this is the value,  $\theta_1 > \theta_0$ , representing a high defect level that is considered unacceptable. If the fraction of defectives found in the lot is such that  $x/n \geq \theta_1$ , then the lot is *definitely* unacceptable.

While the lot is definitely acceptable if  $x/n \leq \theta_0$ , (i.e., the actual fraction (or percent) of defectives is less than the AQL), and while it is definitely unacceptable if  $x/n \geq \theta_1$  (i.e., the actual fraction defectives is greater than the RQL), in between, when  $\theta_0 \leq x/n \leq \theta_1$ , the lot is said to be *barely acceptable*, or of "indifferent quality."

## 3. Types of sampling plans

When the accept/reject decision is to be based on a single sample, the sampling plan is appropriately known as a "single sampling plan." This is the most commonly used plan, but not always the most efficient. With this plan, if  $x < c$ , the lot is accepted; if not, it is rejected. With a double sampling plan, a small initial sample,  $n_1$ , is taken and the number of defectives in the sample,  $x_1$ , determined; the lot is accepted if  $x_1/n_1 \leq \theta_0$  or rejected  $x_1/n_1 \geq \theta_1$ . If the fraction defective is in between, take a second sample of size  $n_2$  and accept or reject the lot on the basis of the two samples. The "multiple sampling plan" is a direct extension of the double sampling plan.

The ideal sampling plan from the consumer's perspective will result in a low probability of accepting a lot for which  $\theta \geq \theta_1$ . For the producer on the other hand, the best sampling plan is one that results in a high probability of accepting a lot for which  $\theta \leq \theta_0$ .

### 22.2.2 Determining a Sampling Plan

Let the random variable  $X$  represent the number of defective items found in a sample of size  $n$ , drawn from a population of size  $N$  whose true but unknown fraction of defectives is  $\theta$ . Clearly, from Chapter 8, we know that  $X$  is a hypergeometric random variable with pdf:

$$f(x) = \frac{\binom{N\theta}{x} \binom{N-N\theta}{n-x}}{\binom{N}{n}} \quad (22.1)$$

The total number of defectives in the lot, of course, is

$$N_D = N\theta; N_D = 0, 1, 2, \dots, N \quad (22.2)$$

At the most basic level, the problem is, *in principle*, that of estimating  $\theta$  from sample data, and testing the hypothesis:

$$H_0 : \theta \leq \theta_0 \quad (22.3)$$

against the alternative

$$H_a : \theta > \theta_0 \quad (22.4)$$

As with all hypothesis tests,  $\alpha$  is the probability of committing a Type I error, in this case, rejecting an acceptable lot; this is the producer's risk. On the other hand,  $\beta$ , the probability of committing a Type II error, not rejecting an unacceptable lot, is the consumer's risk. In practice, sampling plans are designed to balance out these two risks using an approach based on the operating characteristic curve as we now discuss.

### The Operating Characteristic (OC) Curve

Let  $\mathcal{A}$  be the event that the lot under consideration is accepted. From Eq (22.1) as the probability model for the random variable,  $X$ :

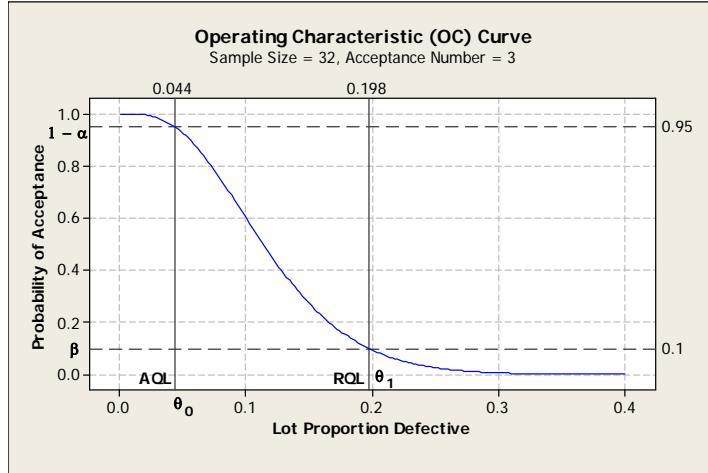
$$\begin{aligned} P(\mathcal{A}) &= P(X \leq c) = \sum_{x=0}^c f(x) \\ &= \sum_{x=0}^c \frac{\binom{N\theta}{x} \binom{N-N\theta}{n-x}}{\binom{N}{n}} \end{aligned} \quad (22.5)$$

Note that  $P(\mathcal{A})$  depends on  $N, \theta$  and  $c$ .

First, let us consider the analysis problem in which  $N$ , the lot size, is known and  $c$ , the acceptance number, is specified, and we simply wish to compute the probability of accepting the lot for various values of  $\theta$ . Under these circumstances, it is customary to rewrite  $P(\mathcal{A})$  as  $P(\mathcal{A}|\theta)$ , so that:

$$P(\mathcal{A}|\theta) = \sum_{x=0}^c f(x|\theta) = \sum_{x=0}^c \frac{\binom{N\theta}{x} \binom{N-N\theta}{n-x}}{\binom{N}{n}} \quad (22.6)$$

We may now note the following about this expression:



**FIGURE 22.1:** OC Curve for a lot size of 1000, sample size of 32 and acceptance number of 3: AQL is the acceptance quality level; RQL is the rejection quality level.

1. Clearly if  $\theta = 0$  (no defective item in the lot), the probability of acceptance is 1; i.e.,

$$P(\mathcal{A}|0) = 1 \quad (22.7)$$

2. As  $\theta$  increases,  $P(\mathcal{A}|\theta)$  decreases; in particular, if  $\theta = 1$ ,

$$P(\mathcal{A}|1) = 0 \quad (22.8)$$

3. A plot of  $P(\mathcal{A}|\theta)$  as a function of  $\theta$  provides information regarding the probability of lot acceptance given the fraction of defectives in the lot.
4. Since  $\theta = N_D/N$  and  $N_D = 0, 1, 2, \dots, N$ , then  $0 < \theta < 1$  can only actually take on values for these discrete values of  $N_D$ , i.e.,  $P(\mathcal{A}|\theta)$  is "defined" only for  $\theta$  values corresponding to  $N_D = 0, 1, 2, \dots, N$ .

Nevertheless, it is customary to connect the valid  $(P(\mathcal{A}|\theta), \theta)$  ordered pairs with a smooth curve, to obtain the operating characteristic curve. An example is shown in Fig 22.1 for the case with  $N = 1000; n = 32; c = 3$ .

What we have presented above is the analysis problem, showing how, given  $N, n$  and  $c$ , we can obtain the acceptance probability  $P(\mathcal{A}|\theta)$  as a function of  $\theta$ , and generate the operating characteristic (OC) curve. In actual fact, the most important use of the OC curve is for generating sampling plans. This is the design problem stated as follows:

Given  $N$ , the lot size, determine from  $P(\mathcal{A}|\theta)$ , feasible values of  $n$  and  $c$  that balance the consumer's and producer's risks.

From a strictly algebraic perspective, determining  $n$  and  $c$  requires generating from Eq (22.6), two equations with these two quantities as the only unknowns, and solving simultaneously. This is achieved as follows: let some value  $p_0$  be selected as the probability of acceptance for lots with defective fraction  $\theta_0$ , and let the probability of acceptance for lots with defective fraction  $\theta_1$  be selected as  $p_1$ . In principle, any arbitrary set of values selected for any of these 4 parameters  $(p_0, \theta_0; p_1, \theta_1)$ , should give us two equations in two unknowns that can be solved for  $n$  and  $c$ . However, it makes more sense to select these parameters judiciously to achieve our objectives. Observe that if these values are selected as follows:

$$p_0 = 1 - \alpha; \quad (22.9)$$

$$p_1 = \beta \quad (22.10)$$

where  $\alpha$  is the producer's risk, and  $\beta$ , the consumer's risk, and if we retain the definitions given above for  $\theta_0$ , the AQL, and  $\theta_1$ , the RQL, then the following two equations:

$$1 - \alpha = \sum_{x=0}^c f(x|\theta_0) = \sum_{x=0}^c \frac{\binom{N\theta_0}{x} \binom{N-N\theta_0}{n-x}}{\binom{N}{n}} \quad (22.11)$$

$$\beta = \sum_{x=0}^c f(x|\theta_1) = \sum_{x=0}^c \frac{\binom{N\theta_1}{x} \binom{N-N\theta_1}{n-x}}{\binom{N}{n}} \quad (22.12)$$

locate two points on the OC curve such that (i) there is a probability  $\alpha$  of rejecting a lot with true defective fraction,  $\theta$ , that is *less* than the AQL,  $\theta_1$ , and (ii) a probability  $\beta$  of accepting (more precisely, not rejecting) a lot with true defective fraction,  $\theta$ , that is *higher* than the RQL. These two equations therefore allow simultaneous consideration of both risks.

Given  $N, \theta_0$  and  $\theta_1$ , the only unknowns in these equations are  $n$  and  $c$ ;  $x$  is an index that runs from 0 to  $c$ . The simultaneous solution of the equations produces the sampling plan. In general, there are no closed form analytical solutions to these equations; they must be solved numerically with the computer. If the specified values for  $\theta_0, \theta_1, \alpha$  and  $\beta$  are reasonable such that a feasible solution of an  $n, c$  pair exists, the obtained solution can then be used to generate the OC curve for the specific problem at hand. Otherwise, the specified parameters will have to be adjusted until a feasible solution can be found.

Thus, to generate a sampling plan, one must specify four parameters: (i)  $\theta_0$ , the acceptable quality level (AQL); (ii)  $\theta_1$ , the rejectable quality level (RQL), along with (iii)  $\alpha$ , the producer's risk, and (iv)  $\beta$ , the consumer's risk. The resulting sampling plan is the pair of values  $n$  and  $c$ , used as follows: the number of samples to take from the lot and test is prescribed as  $n$ ; after testing,  $x$ , the number of defectives found in the sample is compared with  $c$ ; if  $x \leq c$  the lot is accepted; if not the lot is rejected.

### Approximation Techniques

Before presenting an example to illustrate these concepts, we note that the computational effort involved in solving Eqs (22.11) and (22.12) can be reduced significantly by employing well-known approximations to the hypergeometric pdf. First, recall that as  $N \rightarrow \infty$  the hypergeometric pdf tends to the binomial pdf, so that for large  $N$ , Eq (22.6) becomes

$$P(\mathcal{A}|\theta) \approx \sum_{x=0}^c \binom{n}{x} \theta^x (1-\theta)^{n-x} \quad (22.13)$$

which is significantly less burdensome to use, especially for large  $N$ . This is the binomial approximation OC curve.

When the sample size,  $n$ , is large, and hence Eq (22.13) itself becomes tedious, or when the quality assessment is based not on the binary “go/no go” attribute of each tested item but on the number of defects per item, the Poisson alternative to Eq (22.13) is used. This produces the Poisson OC curve. Recall that as  $n \rightarrow \infty$  and  $\theta \rightarrow 0$  but in such a way that  $n\theta = \lambda$ , the binomial pdf tends to the Poisson pdf; then under these conditions, Eq (22.13) becomes

$$P(\mathcal{A}|\theta) \approx \sum_{x=0}^c \frac{(n\theta)^x e^{-n\theta}}{x!} \quad (22.14)$$

#### Example 22.1: SAMPLING PLAN AND OC CURVE FOR ELECTRONIC CHIPS

An incoming lot of 1000 electronic chips is to be evaluated for acceptance on the basis of whether the chips are functioning or not. If the lot contains no more than 40 defectives, it is deemed acceptable; if it contains more than 200 defectives, it is not acceptable. Determine a sampling plan to meet these objectives with an  $\alpha$ -risk of 0.05 and a  $\beta$ -risk of 0.1. Plot the resulting OC curve.

#### Solution:

For this problem, we are given the AQL as  $\theta_0 = 0.04$  and the RQL as  $\theta_1 = 0.2$  along with the standard  $\alpha$  and  $\beta$  risks. The lot size of 1000 is more than large enough to justify using the binomial approximation to determine  $n$  and  $c$ . The MINITAB Quality Tools feature can be used to solve this problem as follows: The sequence: Stat > Quality Tools > Acceptance Sampling by Attribute > opens a dialog box for specifying the problem characteristics. The objective is to Create a Sampling Plan (not Compare user-defined sampling plans); the “measurement type” is Go / no go (defective) (as opposed to number of defects); the “Units for quality levels” is Proportion defective (as opposed to percent defective or defectives per million). The remaining boxes are for the quartet of problem parameters: the AQL, RQL,  $\alpha$ -risk,  $\beta$ -risk, and, in addition, for the lot size,  $N$ . MINITAB also provides options for generating several graphs, of which only the OC curve is selected for this problem. The MINITAB results are:

**Acceptance Sampling by Attributes**

Measurement type: Go/no go

Lot quality in proportion defective

Lot size: 1000

Use binomial distribution to calculate probability of acceptance

|                                |      |
|--------------------------------|------|
| Acceptable Quality Level (AQL) | 0.04 |
| Producer's Risk (Alpha)        | 0.05 |

|  |     |
|--|-----|
| Rejectable Quality Level (RQL or LTPD) | 0.2 |
| Consumer's Risk (Beta)                 | 0.1 |

Generated Plan(s)

|                   |    |
|-------------------|----|
| Sample Size       | 32 |
| Acceptance Number | 3  |

Accept lot if defective items in 32 sampled  $\leq 3$ ; Otherwise reject.

The OC curve is actually the one shown previously Fig 22.1.

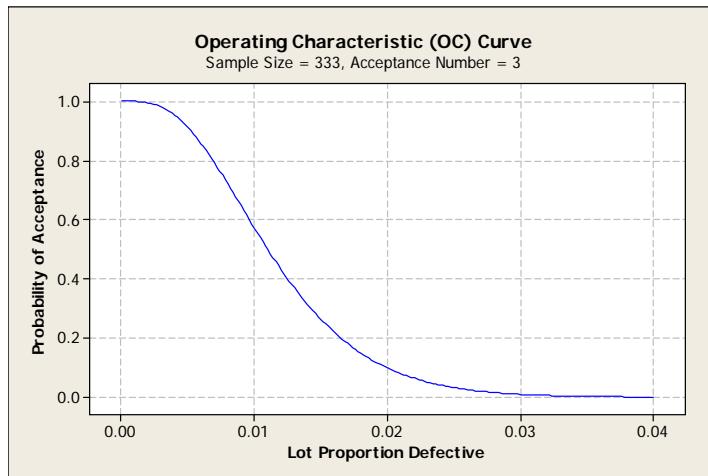
If the AQL and RQL specifications are changed, the sampling plan will change, as will the OC curve. For example, if in Example 22.1 the AQL is changed to 0.004 (only 4 items in 1000 are allowed to be defective for the lot to be acceptable) and the RQL changed to 0.02 (if 20 items or more in 1000 are defective, the lot will be unacceptable), then the sampling plan changes to  $n = 333$  while the acceptance number remains at 3; i.e., 333 samples will be selected for inspection and the lot will be accepted only if the number of defectives found in this sample is 3 or fewer. The resulting OC curve is shown in Fig (22.2) where the reader should pay close attention to the scale on the  $x$ -axis, compared to that in Fig (22.1).

**Characteristics of the OC Curve**

Upon some reflection, we see that the shape of the OC curve is actually indicative of the power of the sampling plan to discriminate between good and bad lots. The steeper the OC curve the better it is at separating good lots from bad; and the larger the sample size,  $n$ , the steeper the OC curve. (Compare, for example, Figs 22.1 and 22.2 on the same scale.)

The shape of the ideal OC curve is a perfect narrow rectangle around the value of the AQL,  $\theta_0$ : every lot with  $\theta = \theta_0$  will be accepted and every lot with  $\theta > \theta_0$  rejected. However, this is unrealistic for many reasons, not least of all being that to obtain such a curve will require almost 100% sampling. The reverse S-shaped curve is more common and more practical. Readers familiar with the theory of signal processing, especially filter design, may recognize the similarity between the OC-curve and the frequency characteristics of filters: the ideal OC curve corresponds to a notch filter while the typical OC curves correspond to low pass, first order filters with time constants of varying magnitude.

Finally, we note that the various discussion of "Power and Sample size"



**FIGURE 22.2:** OC Curve for a lot size of 1000, generated for a sampling plan for an AQL = 0.004 and an RQL = 0.02, leading to a required sample size of 333 and acceptance number of 3. Compare with the OC curve in Fig 22.1.

in Chapter 15 could have been framed in terms of the OC curve; and in fact many textbooks do so.

### Other Considerations

There are other issues associated with acceptance sampling, such as Average Outgoing Quality (AOQ), Average Total Inspection (ATI), and the development of acceptance plans for continuous measures of quality and the concept of the acceptance region; these will not be discussed here, however.

It is important for the reader to recognize that although important from a historical perspective, acceptance sampling is not considered to be very cost-effective as a quality assurance strategy from the perspective of the producer. It does nothing about the process responsible for making the product and has nothing to say about the capability of the process to meet the customer's quality requirements. It is an "after-the-fact," post-production strategy that cannot be the primary tool in the toolbox of a manufacturing enterprise that is serious about producing good quality products.

## 22.3 Process and Quality Control

### 22.3.1 Underlying Philosophy

For a period of time, industrial quality assurance was limited to acceptance sampling: inspecting finished products and removing defective items. From the perspective of the manufacturer, however, to wait until production is complete and then rely on inspection to eliminate poor quality is not a particularly sound strategy. One cannot “inspect” quality into the product.

It is far more efficient that during production, one periodically assesses the product quality via quantitative measures, and analyzes the data appropriately to develop a clear picture of the status of both the process and the product. If the product quality is acceptable, then the process is deemed to be “in control” and no action is necessary; otherwise the process is deemed “out-of-control” and corrective action is taken to restore normal operation. This is the underlying philosophy behind *Process Control* as a quality assurance strategy. There are two primary issues to be addressed in implementing such a strategy:

1. *How should the true value of the process and/or product quality status be determined?* Clearly one cannot sample and analyze all the items produced, not with discrete-parts manufacturing and definitely not in the case of continuous production of, say, specialty chemicals. It is typical to establish a product quality control laboratory where samples taken periodically from the manufacturing process are analyzed; inference can then be drawn from such measurements about the process at-large.
2. *How should corrective action be determined?* As we discuss shortly, this is the central issue that differentiates statistical process control (SPC) from engineering/automatic process control.

### 22.3.2 Statistical Process Control

Statistical Process Control is a popular methodology for implementing the strategy outlined above. It is the application of statistical methods for monitoring process performance over time, enabling the systematic detection of the occurrence of “special cause” events that may require corrective action in order to maintain the process in a state of statistical control.

A process (more precisely, a process variable) is said to be in “statistical control” when the process variable of interest is “statistically stable,” and “on-target.” By statistically stable, we mean that the true process characteristics (typically mean,  $\mu$ , and standard deviation,  $\sigma$ ) are not changing drastically with time; by on-target, we mean that the true process variable mean value

$\mu$ , exactly equals desired target value  $\mu_0$  (or the historical, long-term average value).

At the most fundamental level, therefore, statistical process control involves taking a representative process variable whose value,  $Y$ , is either a direct measure of the product quality of interest, or at least related to it, and assessing whether the observed value,  $y$ , is stable and not significantly different from  $\mu_0$ . Because of inherent variability associated with sampling, and also with the determination of the measured value itself, this problem requires probability and statistics. In particular, observe that one can pose the question: “Is  $y$  significantly different from  $\mu_0$ ?” in the form of the following hypothesis test:

$$\begin{aligned} H_0 : \quad & Y = \mu_0 \\ H_a : \quad & Y \neq \mu_0 \end{aligned} \quad (22.15)$$

a problem we are very familiar with solving, provided an appropriate probability model is available for  $Y$ . In this case, we do not reject the null hypothesis, at the significance level of  $\alpha$ , if  $(Y_L \leq Y \leq Y_U)$ , where the values  $Y_L$  and  $Y_U$  at the rejection boundary are determined from the sampling distribution such that:

$$P(Y_L \leq Y \leq Y_U) = 1 - \alpha \quad (22.16)$$

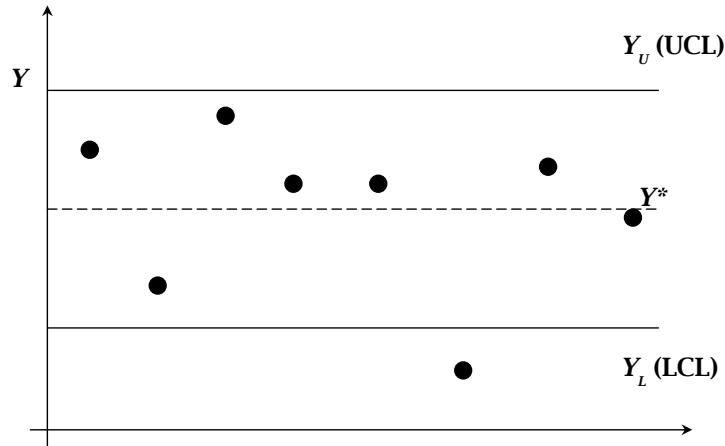
This equation is the foundation of one of an iconic characteristic of SPC; it suggests a convenient graphical technique involving 3 lines:

1. A center line for  $\mu_0$ , the desired target for the random variable,  $Y$ ;
2. An “Upper Control Limit” (UCL) line for  $Y_U$ ; and
3. A “Lower Control Limit” (LCL) line for  $Y_L$

on which each acquired value of  $Y$  is plotted. Observe then that a point falling outside of these limits signals the rejection of the null hypothesis in favor of the alternative, at the significance level of  $\alpha$ , indicating an “out-of-control” status. A generic SPC chart of this sort is shown in Fig 22.3 where the sixth data point is out of limits. Points within the control limits are said to show variability attributable to “common cause” effects; “special cause” variability is considered responsible for points falling outside the limits.

In traditional SPC, when an “out-of-control” situation is detected, the recommended corrective action is to “find and eliminate” the problem, the practical implementation of which is obviously process-specific so that the instruction cannot be more specific than this. But in the discrete-parts manufacturing industry, there is significant cost associated with finding and correcting problems. There is therefore significant incentive to minimize false “out-of-control” alarms.

Finally, before beginning the discussion of specific charts, we note that the nature of the particular quantity  $Y$  that is of interest in any particular case



**FIGURE 22.3:** A generic SPC chart for the generic process variable  $Y$  indicating a sixth data point that is out of limits.

clearly determines the probability model underlying the chart, which in turn determined how  $Y_U$  and  $Y_L$  are determined. The ensuing discussion of various SPC charts is from this perspective.

### 22.3.3 Basic Control Charts

Control charts are graphical (visual) means of monitoring process characteristics. They typically consist of two plots: one for monitoring the “mean” value of the process variable in question; the other for monitoring the “variability,” although the chart for the mean customarily receives more attention. These charts are nothing but graphical means of carrying out the hypotheses tests:  $H_0$ : Process Mean = Target; Process Variability = Constant, versus the alternative:  $H_a$ : Process mean  $\neq$  Target; and/or Process Variability  $\neq$  Constant. In practice, these tests are implemented in real-time by adding each new set of process/product data as they become available. Modern implementations involve displays on computer screens that are updated at fixed intervals of time, with alarms sounding whenever an “alarm-worthy” event occurs.

It is important to stress that the control limits indicated in Fig 22.3 are not *specification limits*; these control limits strictly arise from the sampling distribution of the process variable,  $Y$ , and are indicative of typical variability intrinsic to the process. The control limits enable us determine if observed variability is in line with what is typical. In the language of the quality movement, these control limits therefore constitute the “voice of the process.” Specification limits on the other hand have nothing to do with the process; they are specified by the customer, independent of the process, and therefore constitute what is known as the “voice of the customer.”

A few of the various charts that exist for various process and product variables and attributes are now discussed.

### The Shewhart Xbar Chart

By far the oldest, most popular and most recognizable control chart is the Shewhart chart, named for Walter A. Shewhart (1891–1967), the Bell Labs physicist and engineer credited with pioneering industrial statistical quality control. In its most basic form, it is a chart used to track the sample mean,  $\bar{X}$ , of a process or product variable: for example, the mean outer diameter of ball bearings; the mean length of 6-inch nails; the mean liquid volume of 12-ounce cans of soda; the mean Mooney viscosity of several samples of an elastomer, etc. The generic variable  $Y$  in this case is  $\bar{X}$ , the sample mean of the process measurements.

The data requirement is as follows: a random sample,  $X_1, X_2, \dots, X_n$  is obtained from the process in question, from which the average,  $\bar{X}$ , and standard deviation,  $S_{\bar{X}}$ , are computed. The probability model underlying the Shewhart chart is the gaussian distribution, justified as follows. There are many instances where the variable of interest,  $X$ , is itself approximately normally distributed, in which case  $\bar{X} \sim N(\mu_0, \sigma_{\bar{X}}^2)$ ; but even when  $X$  is not normally distributed, for most random variables,  $N(\mu_0, \sigma_{\bar{X}}^2)$  is a reasonable approximate distribution for  $\bar{X}$ , given a large enough sample (as a result of the Central Limit Theorem).

With this sampling distribution for  $\bar{X}$ , we are able to compute the following probability:

$$P(-3\sigma_{\bar{X}} < \bar{X} - \mu_0 < 3\sigma_{\bar{X}}) = 0.9974 \quad (22.17)$$

providing the characteristic components of the Shewhart chart: the control limits are  $\pm 3\sigma_{\bar{X}}$  to each side of the target value  $\mu_0$  on the center line; and the confidence level is  $(1-\alpha) \times 100\% = 99.7\%$ . The bounds are therefore commonly known as “3-sigma limits.” The  $\alpha$ -risk of false “out-of-control” alarms is thus very low at 0.003. An example follows.

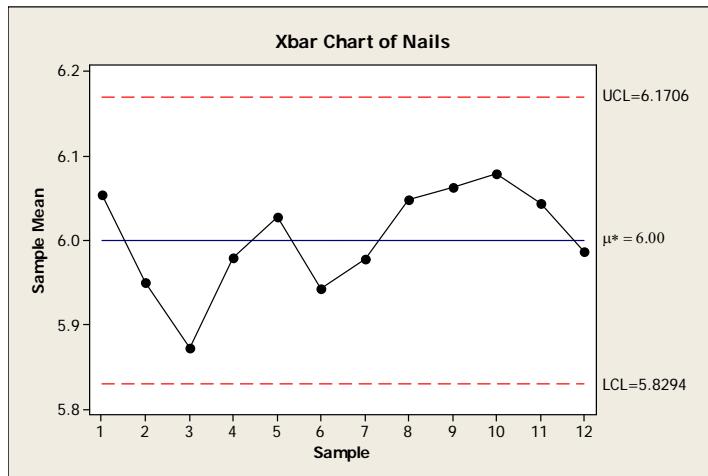
#### **Example 22.2: X-BAR CONTROL CHART FOR 6-INCH NAILS**

Every five minutes, a random sample of 3 six-inch nails is selected from a manufactured lot and measured for conformance to the specification. The data in Table 22.1 is a record of the measurements determined over the first hour of a shift. Obtain an X-bar chart and identify whether or not the manufacturing process is “in-control”.

#### **Solution:**

The points to be plotted are the averages of the three samples corresponding to each sample time; the center line is the target specification of 6 inches. To obtain the control limits, however, observe that we have not been given the process standard deviation. This is obtained from the data set itself, assuming that the process is “in-control.”

Computer programs such as MINITAB can be used to obtain



**FIGURE 22.4:** The X-bar chart for the average length measurements for 6-inch nails determined from samples of three measurements obtained every 5 mins.

the desired X-bar chart. Upon entering the data into a worksheet, in MINITAB, the sequence **Stat>Control Charts>Variables Charts for Subgroups> X-bar>** opens a self-explanatory dialog where the problem characteristics are entered. The result is the chart shown in Fig 22.4. Observe that the entire collection of data, the twelve average values, are all within the control limits, implying that the process appears to be “in control.”

### The S-Chart

The objective of the original Shewhart chart is to determine the status of the process with respect to  $\bar{X}$ , the mean value of the process/product variable of interest. But this is not the only process/product characteristic of interest. The average,  $\bar{X}$ , may remain on target while the variability may have changed. There are cases of practical importance where the variability is the primary variable of interest, especially when we are concerned with detecting if the process variability has changed significantly.

Under these circumstances, the variable of interest,  $Y$ , is now  $S_X$ , the sample standard deviation, determined from the same random sample of size  $n$  used to obtain  $\bar{X}$ . The probability model is obtained from the fact that, for a sample size of  $n$ ,

$$S_X = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}} \quad (22.18)$$

**TABLE 22.1:** Measured length of samples of 6-inch nails in a manufacturing process

| Length<br>(in) | Sample<br># | Length<br>(in) | Sample<br># | Length<br>(in) | Sample<br># |
|----------------|-------------|----------------|-------------|----------------|-------------|
| 6.05           | 1           | 6.01           | 5           | 6.06           | 9           |
| 6.10           | 1           | 6.17           | 5           | 6.06           | 9           |
| 6.01           | 1           | 5.90           | 5           | 6.07           | 9           |
| 5.79           | 2           | 5.86           | 6           | 6.09           | 10          |
| 5.92           | 2           | 6.03           | 6           | 6.07           | 10          |
| 6.14           | 2           | 5.93           | 6           | 6.07           | 10          |
| 5.86           | 3           | 6.17           | 7           | 6.00           | 11          |
| 5.90           | 3           | 5.81           | 7           | 6.10           | 11          |
| 5.86           | 3           | 5.95           | 7           | 6.03           | 11          |
| 5.88           | 4           | 6.10           | 8           | 5.98           | 12          |
| 5.98           | 4           | 6.09           | 8           | 6.01           | 12          |
| 6.08           | 4           | 5.95           | 8           | 5.97           | 12          |

and, from previous discussions in Chapters 14 and 15, we know that

$$\frac{(n-1)S_X^2}{\sigma_X^2} \sim \chi^2(n-1) \quad (22.19)$$

where  $\sigma_X^2$  is the inherent process variance. It can be shown, first, that

$$E(S_X) = c(n)\sigma_X \quad (22.20)$$

where the sample-size-dependent constant  $c(n)$  is given by

$$c(n) = \left( \sqrt{\frac{2}{n-1}} \right) \frac{\Gamma(n/2)}{\Gamma(\frac{n-1}{2})} \quad (22.21)$$

so that  $S_X/c(n)$  is unbiased for  $\sigma_X$ , providing us with an expression for the “center line.”

Obtaining the control limits for  $S_X$  requires a measure of the variability associated with estimating  $\sigma_X$  with  $S_X/c(n)$ . In this regard, it can be shown that

$$\sigma_{S_X}^2 = \sigma_X^2 [1 - c^2(n)] \quad (22.22)$$

We may now combine all these expressions to obtain that

$$P[-3\sigma_{S_X} < S_X - c(n)\sigma_X < 3\sigma_{S_X}] \approx 0.99 \quad (22.23)$$

so that:

$$P[(c(n)\sigma_X - 3\sigma_{S_X}) < S_X < (c(n)\sigma_X + 3\sigma_{S_X})] \approx 0.99 \quad (22.24)$$

from which the control limits for the  $S$ -chart are obtained as:

$$UCL = \sigma_X \left[ c(n) + 3\sqrt{1 - c^2(n)} \right] \quad (22.25)$$

$$LCL = \sigma_X \left[ c(n) - 3\sqrt{1 - c^2(n)} \right] \quad (22.26)$$

Thus, when the process is “in-control” with respect to variability, observed values of  $S_X$  will lie within the 3-sigma bounds indicated by Eqs (22.25) and (22.26) approximately 99% of the time. Values outside of these bounds signal a special-cause event, at this confidence level.

And now, here are some practical considerations: First,  $\sigma_X$  is usually not available; it is typical to estimate it from process data. For example, from several samples with standard deviations,  $S_1, S_2, \dots, S_j$ , the average,

$$\bar{S} = \frac{\sum_{i=1}^j S_i}{j} \quad (22.27)$$

is used to estimate  $\sigma_X$  as  $\bar{S}/c(n)$  where  $n$  (not the same as  $j$ ) is the total number of data points employed to determine  $S_X$ . Under these circumstances, Eqs (22.25) and (22.26) become:

$$UCL = \bar{S} \left[ 1 + 3\sqrt{\left( \frac{1}{c^2(n)} - 1 \right)} \right] \quad (22.28)$$

$$LCL = \bar{S} \left[ 1 - 3\sqrt{\left( \frac{1}{c^2(n)} - 1 \right)} \right] \quad (22.29)$$

Whenever the computed LCL is negative, it is set to zero for the obvious reason that standard deviation is non-negative. Finally, these somewhat intimidating-looking computations are routinely carried out by computer programs. For example, the  $S$ -chart for the data used in Example 22.2 is shown here in Fig 22.5. It is obtained from MINITAB using the sequence: **Stat > Control Charts > Variables Charts for Subgroups > S >**; it shows that the process variability is itself reasonably steady.

It is typical to combine the X-bar and  $S$  charts to obtain the “Xbar-S” chart. This composite chart allows one to confirm that the process is *both* on-target (indicated by the Xbar component) and stable (indicated by the  $S$  component). It is possible for the process to be stable and on-target (the preferred state); stable but not on-target; not stable and not on-target; and less likely (but not impossible), on-target but stable. The combination “Xbar-S” chart allows the determination of which of these four possible states best describes the process.

### Variations to the Xbar-S Chart: Xbar-R, and I & MR Charts

Sometimes the process data sample size is not large enough to provide reasonable estimates of the standard deviation,  $S$ . In such cases, the sample

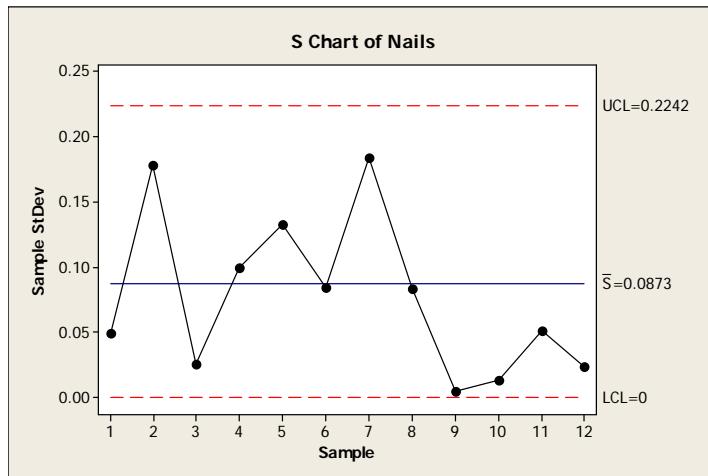


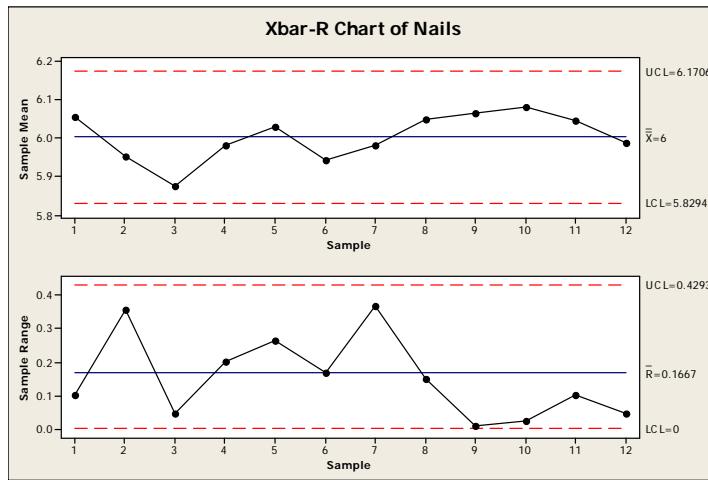
FIGURE 22.5: The *S*-chart for the 6-inch nails process data of Example 22.2.

range,  $R$  (the difference between the lowest and the highest ranked observations in the sample) is used as a measure of process variability. This gives rise to the  $R$ -chart by itself, or the  $X\bar{}$ - $R$  chart when combined with the  $X\bar{}$  chart. The same principles discussed previously apply: the chart is based on a probability model for the random variable,  $R$ ; its expected value and its theoretical variance are used to obtain the control limits.

In fact, because the data for the process in Example 22.2 involves samples of size  $n = 3$ , it is questionable whether this sample size is sufficient for obtaining reliable estimates of sample standard deviation,  $\sigma$ . When  $n < 8$ , it is usually recommended to use the  $R$  chart instead.

The combination  $X\bar{}$ - $R$  chart for the “Nails” data of Example 22.2 is shown here in Fig 22.6. It is obtained from MINITAB using the sequence: Stat > Control Charts > Variables Charts for Subgroups >  $X\bar{}$ - $R$  >. The most important points to note are: (i) the chart still indicates that the process variability is reasonably steady; however, (ii) the nominal value for  $R$  is almost twice that for  $S$  (this is expected, given the definition of the range  $R$  and its relation to the standard deviation,  $S$ ); by the same token, the control limits are also wider (approximately double); nevertheless, (iii) the general characteristics of the  $R$  component of the chart is not much different from that of the  $S$ -chart obtained earlier and shown in Fig 22.5. Thus, in this case, the  $S$  and  $R$  variables show virtually the same characteristics.

In many cases, especially common in chemical processes, only individual measurements are available at each sampling time. Under these circumstances, with sample size  $n = 1$ , one can definitely plot the individual measurements against the control limits, so that this time, the variable  $Y$  is now the actual



**FIGURE 22.6:** The combination Xbar-R chart for the 6-inch nails process data of Example 22.2.

process measurement  $X$ , not the average; but with no other means available for estimating intrinsic variability, it is customary to use the moving range, defined as:

$$MR_i = |X_i - X_{i-1}| \quad (22.30)$$

the difference between consecutive observations, as a measure of the variability. This combination gives rise to the “I and MR” chart (Individual and Moving Range). The components of this chart are also determined using the same principles as before: the individual samples are assumed to come from a gaussian distribution, providing the probability model for the “I” chart, from which the control limits are obtained, given an estimate of process variability,  $\sigma$ . Upon assuming that individual observations are mutually independent, the expected value and theoretical variance of the moving range are used to obtain the control limits for the MR chart. We shall return shortly to the issue of the independence assumption. For now we note once more that the required computations are easily carried out with computer programs. The following example illustrates the I and MR chart for a polymer process.

**Example 22.3: CONTROL CHART FOR ELASTOMER PROCESS**

Ogunnaike and Ray, (1994)<sup>1</sup> presented in Chapter 28, hourly lab measurements of Mooney viscosity obtained for a commercial elastomer manufactured in a continuous process. The data set is reproduced here in Table 22.2. If the desired target Mooney viscosity value for this product is 50.0, determine whether or not the process is stable and on target.

<sup>1</sup>B.A. Ogunnaike, and W.H. Ray, (1994). *Process Dynamics, Modeling and Control*, Oxford, NY.

**TABLE 22.2:** Hourly  
Mooney viscosity data

| Time Sequence<br>(in hours) | Mooney<br>Viscosity |
|-----------------------------|---------------------|
| 1                           | 49.8                |
| 2                           | 50.1                |
| 3                           | 51.1                |
| 4                           | 49.3                |
| 5                           | 49.9                |
| 6                           | 51.1                |
| 7                           | 49.9                |
| 8                           | 49.8                |
| 9                           | 49.7                |
| 10                          | 50.8                |
| 11                          | 50.7                |
| 12                          | 50.5                |
| 13                          | 50.0                |
| 14                          | 50.3                |
| 15                          | 49.8                |
| 16                          | 50.8                |
| 17                          | 48.7                |
| 18                          | 50.4                |
| 19                          | 50.8                |
| 20                          | 49.6                |
| 21                          | 49.9                |
| 22                          | 49.7                |
| 23                          | 49.5                |
| 24                          | 50.5                |
| 25                          | 50.8                |

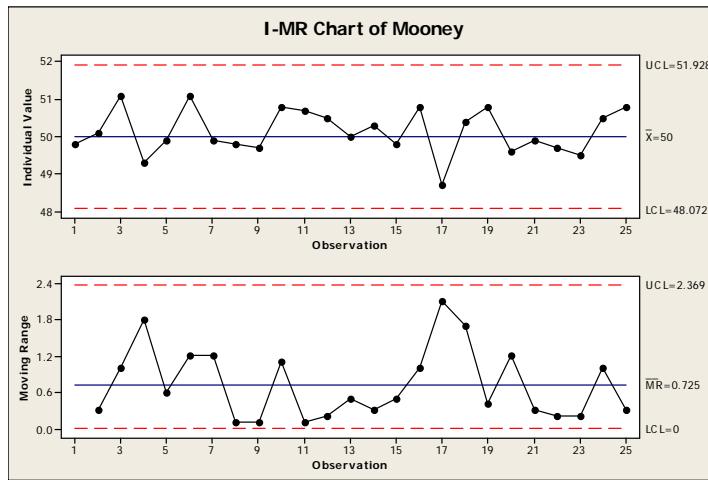


FIGURE 22.7: The combination I-MR chart for the Mooney viscosity data.

### Solution:

Because we only have individual observations at each sampling time, this calls for an I and MR chart. Such a chart is obtained from MINITAB using the sequence: Stat > Control Charts > Variables Charts for Individual > I-MR >. The result is shown in Fig 22.7, which indicates that the process is in “statistical control.”

### The P-Chart

When the characteristic of interest is the proportion of defective items in a sample, the appropriate chart is known as the *P*-chart. If  $X$  is the random variable representing the number of defective items in a random sample of size  $n$ , then we know from Chapter 8 that  $X$  possesses a binomial distribution, in this case,

$$f(x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x} \quad (22.31)$$

where  $\theta$  is the true but unknown population proportion of defectives. The maximum likelihood estimator,

$$P = \frac{X}{n} \quad (22.32)$$

is unbiased for  $\theta$ . From the characteristics of the binomial random variable, we know that

$$E(X) = n\theta \quad (22.33)$$

$$\sigma_X^2 = n\theta(1 - \theta) \quad (22.34)$$

so that

$$E(P) = \theta \quad (22.35)$$

$$\sigma_P^2 = \frac{\theta(1-\theta)}{n} \quad (22.36)$$

From data consisting of  $k$  separate samples of size  $n$  each, yielding  $k$  actual proportions of defectives,  $p_1, p_2, \dots, p_k$ ,  $\theta$  is estimated as  $\bar{p}$  defined as:

$$\bar{p} = \frac{\sum_{i=1}^k p_i}{k} \quad (22.37)$$

with the associated standard deviation,

$$\hat{\sigma}_p = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \quad (22.38)$$

These results can be used to construct the  $P$ -chart to monitor the proportion of defectives in a manufacturing process. The center line is the traditional long term average  $\bar{p}$ , and the 3-sigma control limits are:

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \quad (22.39)$$

$$LCL = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \quad (22.40)$$

Once again, negative values of LCL are set to zero. The following example illustrates the  $P$ -chart.

**Example 22.4: IDENTIFYING “SPECIAL CAUSE” IN MECHANICAL PENCIL PRODUCTION**

A mechanical pencil manufacturer takes a sample of 10 every shift and tests the lead release mechanism. The pencil is marked defective if the lead release mechanism does not function as prescribed. Table 22.3 contains the results from 10 consecutive shifts during a certain week in the summer; it shows the sample size, the number of defective pencils identified and the proportion defective. Obtain a control chart for the data and assess whether or not the manufacturing process is “in control.”

**Solution:**

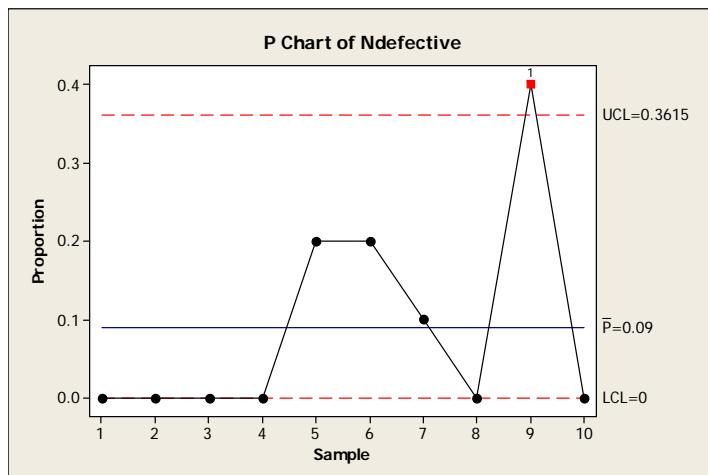
The  $P$ -chart obtained from these data using MINITAB (sequence: **Stat > Control Charts > Attributes Charts > P >**) is shown in Fig 22.8 where one point, the entry from the 9<sup>th</sup> shift, falls outside of the UCL. The MINITAB output is as follows:

**Test Results for P Chart of Ndefective**

TEST 1. One point more than 3.00 standard deviations from center

**TABLE 22.3:** Number and proportion of defective mechanical pencils

| Shift | Sample Size | Number defective | Proportion defective |
|-------|-------------|------------------|----------------------|
| 1     | 10          | 0                | 0.0                  |
| 2     | 10          | 0                | 0.0                  |
| 3     | 10          | 0                | 0.0                  |
| 4     | 10          | 0                | 0.0                  |
| 5     | 10          | 2                | 0.2                  |
| 6     | 10          | 2                | 0.2                  |
| 7     | 10          | 1                | 0.1                  |
| 8     | 10          | 0                | 0.0                  |
| 9     | 10          | 4                | 0.4                  |
| 10    | 10          | 0                | 0.0                  |



**FIGURE 22.8:** P-chart for the data on defective mechanical pencils: note the 9<sup>th</sup> observation that is outside the UCL.

line.

**Test Failed at points: 9**

Upon further review, it was discovered that during shift 9 (Friday morning) was when a set of new high school summer interns were being trained on how to run parts of the manufacturing process; the mistakes made were promptly rectified and the process returned to normal by the end of shift 10.

### The C-Chart

When the process/product characteristic of interest is the *number* of defects per item (for example, the number of inclusions on a glass sheet of given area, as introduced in Chapter 1 and revisited several times in ensuing chapters), the appropriate chart is the C-chart. This chart, like the others, is developed on the basis of the appropriate probability model, which in this case, is the Poisson model. This is because  $X$ , the random variable representing the number of defects per item is Poisson-distributed, with pdf

$$f(x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad (22.41)$$

(The tradition is to represent this attribute by “C” for count). Observe from what we know about the Poisson random variable that

$$\mu_C = \lambda \quad (22.42)$$

$$\sigma_C^2 = \lambda \quad (22.43)$$

Thus, once again, from a random sample,  $X_1, X_2, \dots, X_k$ , which represents the number of defects found on  $k$  separate items, one can estimate  $\lambda$  from the sample average:

$$\hat{\lambda} = \frac{\sum_{i=1}^k X_i}{k} \quad (22.44)$$

from where the 3-sigma control limits are obtained as:

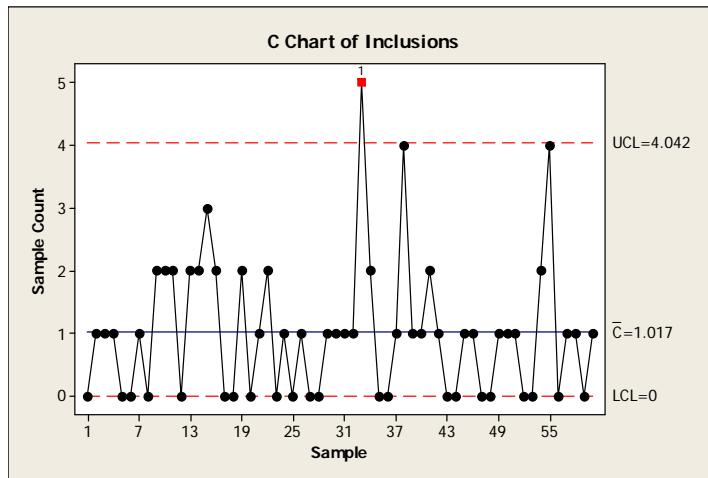
$$UCL = \hat{\lambda} + 3\sqrt{\hat{\lambda}} \quad (22.45)$$

$$LCL = \hat{\lambda} - 3\sqrt{\hat{\lambda}} \quad (22.46)$$

setting LCL = 0 in place of negative values.

An example C-Chart for the inclusions data introduced in Chapter 1 (Table 1.2), is shown in Fig 22.9, obtained from MINITAB using the sequence **Stat > Control Charts > Attributes Charts > C >**. The lone observation of 5 inclusions (in the 33<sup>rd</sup> sample) is flagged as out of limit; otherwise, the process seems to be operating in control, with an average number of inclusions of approximately 1, and an upper limit of 4 (see Eq (22.45) above).

If we recall the discussion in Chapter 15, especially Example 15.15, we note



**FIGURE 22.9:** C-chart for the inclusions data presented in Chapter 1, Table 1.2, and discussed in subsequent chapters: note the 33<sup>rd</sup> observation that is outside the UCL, otherwise, the process appears to be operating in statistical control

that this C-chart is nothing but a visual, graphical version of the hypothesis test carried out in that example. We concluded then that the process was on target (at that time, at the 95% confidence level); we reach the same conclusion with this chart, at the 99% confidence level.

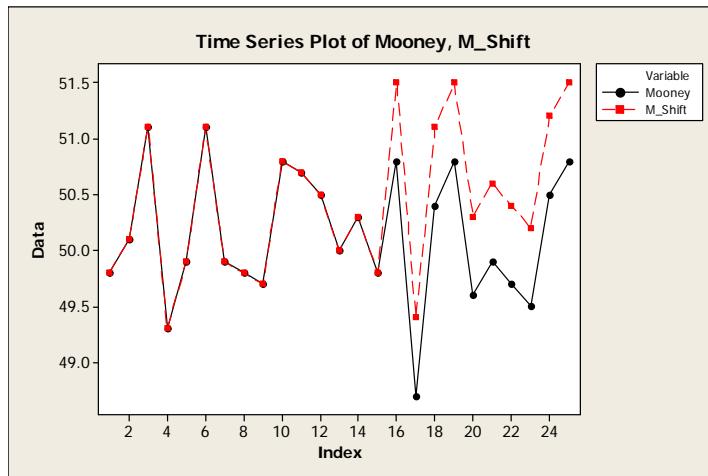
#### 22.3.4 Enhancements

##### Motivation

Basic SPC charts, as originally conceived, needed enhancing for several reasons, the three most important being:

1. Sensitivity to small shifts;
2. Serial correlation;
3. Multivariate data

It can be truly challenging for standard SPC charts to detect small changes. This is because of the very low  $\alpha$ -risk ( $\alpha = 0.003$  compared to  $\alpha = 0.05$  used for hypothesis tests) chosen to prevent too many false out-of-control alarms. The natural consequence is that the  $\beta$ -risk of failing to identify an out-of-control situation increases. To illustrate, consider the Mooney viscosity data shown in Table 22.2; if a step increase in Mooney viscosity of 0.7 occurs after sample 15 and persists, a sequence plot of both the original data and the shifted data is shown in Fig 22.10, where the shift is clear. However, an I-Chart for the shifted data, shown in Fig 22.11, even after specifying the population standard



**FIGURE 22.10:** Time series plot of the original Mooney viscosity data of Fig 22.7 and Table 22.2, and of the shifted version showing a step increase of 0.7 after sample 15.

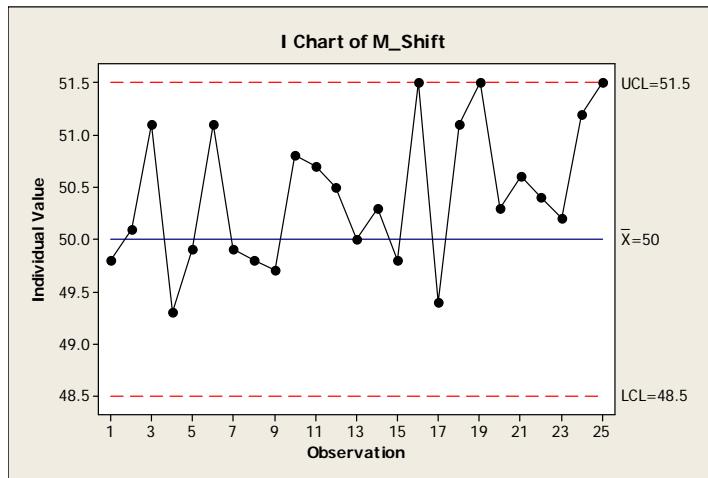
deviation as  $\sigma = 0.5$  (less than the value of approximately 0.62 used for the original data in Fig 22.7), is unable to detect the shift.

Techniques employed to improve the sensitivity to small changes, include the “Western Electric Rules<sup>2</sup>,” the CUSUM (Cumulative Sum) chart, and the EWMA (Exponentially Weighted Moving Average) chart, which will all be discussed shortly.

The issue of serial correlation is often a key characteristic of industrial chemical processes where process dynamics are significant. Classical SPC assumes no serial correlation in process data, and that mean shifts occur only due to infrequent “special causes.” The most direct way to handle this type of process variability is “Engineering/Automatic Process Control.” We will also introduce this briefly.

Finally, industrial processes are intrinsically multivariable so that the data used to track process performance come from several process variables, sometimes numbering in the hundreds. These process measurements are such that if, for example,  $\mathbf{y}_1 = \{y_{11}, y_{12}, \dots, y_{1n}\}$  represents the sequence of observations for one variable, say reactor temperature, and there are others just like it,  $\mathbf{y}_k = \{y_{k1}, y_{k2}, \dots, y_{kn}\}$ , sequences from other variables,  $k = 2, 3, \dots, m$ , (say reactor pressure, agitator amps, catalyst flow rate, etc.) then the sequences  $\mathbf{y}_j$  and  $\mathbf{y}_\ell$  with  $j \neq \ell$  are often highly correlated. Besides this, it is also impossible to visualize the entire data collection properly with the usual single, individual variable SPC charts. Special multivariate techniques for dealing with this type of process variability will be discussed in Chapter 23.

<sup>2</sup>Western Electric Company (1956), *Statistical Quality Control Handbook*. (1<sup>st</sup> Edition.), Indianapolis, Indiana.



**FIGURE 22.11:** I-chart for the shifted Mooney viscosity data. Even with  $\sigma = 0.5$ , it is not sensitive enough to detect the step change of 0.7 introduced after sample 15.

### Western Electric Rules

The earliest enhancement to the Shewhart chart came in the form of what is known as the *Western Electric Rules*. With the standard control limits set at  $\pm 3\sigma$  from the center line, the following is a version of these rules (with the original Shewhart condition as the first rule).

A special event is triggered when:

1. One point falls outside the  $\pm 3\sigma$  limits; or
2. Two of 3 consecutive points fall outside the  $\pm 2\sigma$  limits; or
3. Four of 5 consecutive points fall outside the  $\pm 1\sigma$  limits; or
4. Eight consecutive points fall on either side of the center line.

These additional rules derive from event probabilities for random samples drawn from gaussian distributions and have been known to improve the standard chart sensitivity to small changes. Almost all statistical software packages include these additional detection rules as user-selected options.

### CUSUM Charts

Instead of plotting individual observations  $X_i$ , consider a strategy based on  $S_i$ , the cumulative sum of deviations from desired target, defined as:

$$S_n = \sum_{i=1}^n (X_i - \mu_0) \quad (22.47)$$

This quantity has the following distinctive characteristics: (i) random variations around the target manifest as a “random walk,” an accumulation of small, zero mean, random errors; on the other hand, (ii) if there is a shift in mean value—no matter how slight—and it persists, this event will eventually translate to a noticeable change in character, an upward trend for a positive shift, or a downward trend for a negative shift. As a result of the persistent accumulation, the slope of these trends will be related to the magnitude of the change.

CUSUM charts, of which there are two types, are based on the probabilistic characterization of the random variable,  $S_n$ . The one-sided CUSUM charts are plotted in pairs: an upper CUSUM to detect positive shifts (an increase in the process variable value), and the lower CUSUM to detect negative shifts. The control limits, UCL and LCL, are determined in the usual fashion on the basis of the appropriate sampling distribution (details not considered here). This version is usually preferred because it is easier to construct and to interpret. It is also possible to obtain a single two-sided CUSUM chart. Such a chart uses a so-called V-mask instead of the typical 3-sigma control limits. While the intended scope of this discussion does not extend beyond this brief overview, additional details regarding the CUSUM chart are available, for example, in Page (1961)<sup>3</sup>, and Lucas (1976)<sup>4</sup>.

Fig 22.12 shows the two one-sided CUSUM charts corresponding directly to the I-Chart of Fig 22.11, with the standard deviation specified as 0.5, and the target as 50. (The chart is obtained from MINITAB with the sequence: Stat > Control Charts > Time-Weighted Charts > CUSUM >). The upper CUSUM for detecting positive shifts is represented with dots; the lower CUSUM with diamonds, and the non-conforming data with squares. Note that very little activity is manifested in the lower CUSUM. This is in contrast to the upper CUSUM where the influence of the introduced step change is identified after sample 18, barely three samples after its introduction. Where the I-Chart based on individual observations is insensitive to such small changes, the amplification effect of the error accumulation implied in Eq 22.47 has made this early detection possible.

For the sake of comparison, Fig 22.13 shows the corresponding one-sided CUSUM charts for the original Mooney viscosity data, using the same characteristics as the CUSUM charts in Fig 22.12; no point is identified as non-conforming, consistent with the earlier analysis of the original data.

### EWMA Charts

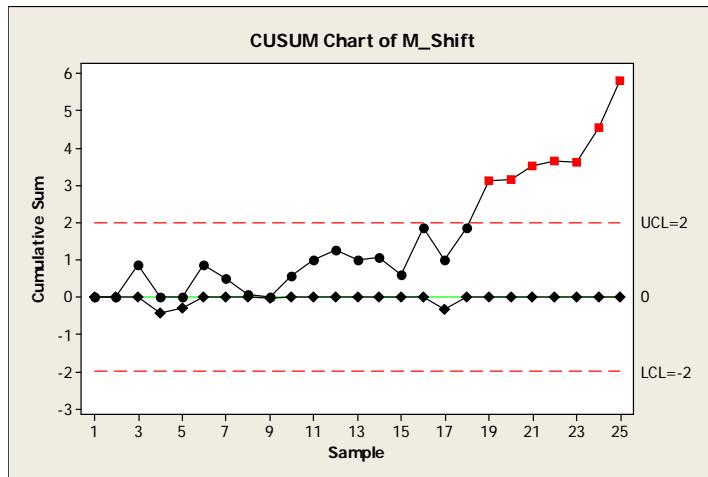
Rather than plot the individual observation  $X_i$ , or the cumulative sum shown in Eq (22.47), consider instead the following variable,  $Z_i$ , defined by:

$$Z_i = wX_i + (1 - w)Z_{i-1}; 0 \leq w \leq 1 \quad (22.48)$$

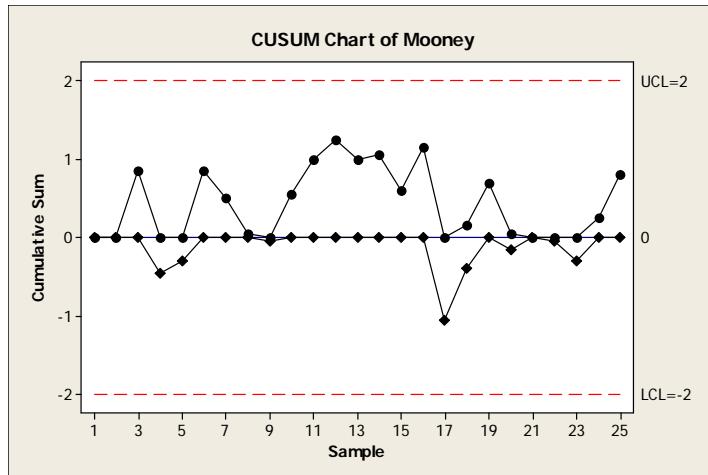
---

<sup>3</sup>E.S. Page (1961). “Cumulative Sum Charts,” *Technometrics*, 3, 1-9.

<sup>4</sup>J.M. Lucas (1976). “The Design and Use of V-Mask Control Schemes,” *Journal of Quality Technology*, 8, 1-12.



**FIGURE 22.12:** Two one-sided CUSUM charts for the shifted Mooney viscosity data. The upper chart uses dots; the lower chart uses diamonds; the non-conforming points are represented with the squares. With the same  $\sigma = 0.5$ , the step change of 0.7 introduced after sample 15 is identified after sample 18. Compare with the I-Chart in Fig 22.11.



**FIGURE 22.13:** Two one-sided CUSUM charts for the original Mooney viscosity data using the same characteristics as those in Fig 22.12. The upper chart uses dots; the lower chart uses diamonds; there are no non-conforming points.

a “filtered” value of  $X_i$ . By recursive substitution, we obtain  $Z_i$  as

$$Z_i = wX_i + w(1-w)X_{i-1} + w(1-w)^2X_{i-2} + \cdots + w(1-w)^{i-1}X_1 + w(1-w)^iX_0 \quad (22.49)$$

an exponentially weighted moving average of the past values of  $X$ . Therefore  $Z_i$  is simply a smoother version of the original data sequence.

Charts based on  $Z_i$  are known as exponentially-weighted-moving-average (EWMA) charts because of Eq (22.49). The premise is that by choosing the weights  $w$  appropriately, small shifts in  $X$  can be detected fairly rapidly in the resulting  $Z$  sequence. The performance of EWMA charts therefore depends on the values chosen for the design parameter,  $w$ ; and for certain choices, these charts are related to Shewhart and CUSUM charts:

1. When  $w = 1$ , the EWMA chart is identical to the basic Shewhart chart;
2. For any other value, the EWMA chart provides a compromise between the Shewhart chart with “no memory” of past data, and the CUSUM chart with infinite memory (the entire data history being “carried along” in the cumulative sum). The EWMA employs  $w(1-w)^{i-k}$  as a “forgetting factor” which determines by how much  $X_k$  influences  $Z_i$ ,  $i > k$ , in such a way that data farther from current time  $i$  exert less influence on  $Z_i$  than more current data.
3. The smaller the value of  $w$ , the greater the influence of historical data, and the further away from the basic Shewhart chart—and the closer to the CUSUM chart—the EWMA chart becomes.
4. It can be shown that specifically for  $w = 0.4$ , the EWMA closely approximates the Shewhart chart in combination with Western Electric rules.

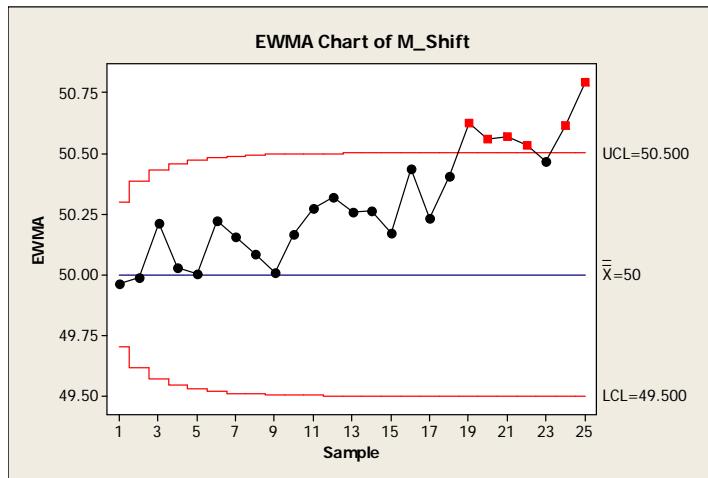
As with other charts, the characteristics of the EWMA chart, especially the control limits, are determined from the sampling distribution of the random variable,  $Z_i$ . For additional details, the reader is referred to Lucas and Saccucci, (1990)<sup>5</sup>.

For purposes of illustration, Fig 22.14 shows the EWMA chart for the shifted Mooney viscosity data, corresponding directly to the I-Chart of Fig 22.11. The value chosen for the design parameter is  $w = 0.2$ ; the standard deviation is specified as 0.5, and the target as 50 (as in the I-Chart and the CUSUM chart). The EWMA chart is obtained from MINITAB with the sequence: Stat > Control Charts > Time-Weighted Charts > EWMA >.

Note the distinctive staircase shape of the control limits: tighter at the beginning, becoming wider as more data become incorporated into the exponentially weighted moving average. As with the CUSUM chart, the shift is detected after sample 18. For comparison, the corresponding EWMA chart

---

<sup>5</sup>Lucas, J.M., and M.S. Saccucci (1990). “Exponentially Weighted Moving Average Control Schemes: Properties and Enhancements,” *Technometrics*, 32, 1-29.



**FIGURE 22.14:** EWMA chart for the shifted Mooney viscosity data, with  $w = 0.2$ . Note the staircase shape of the control limits for the earlier data points. With the same  $\sigma = 0.5$ , the step change of 0.7 introduced after sample 15 is detected after sample 18. The non-conforming points are represented with the squares. Compare with the I-Chart in Fig 22.11 and the CUSUM charts in Fig 22.12.

for the original Monney viscosity data is shown in Fig 22.15; as expected, no non-conforming points are identified.

## 22.4 Chemical Process Control

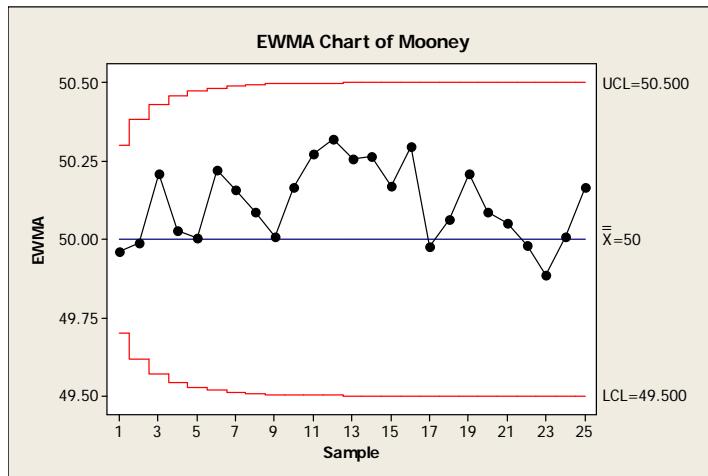
### 22.4.1 Preliminary Considerations

For chemical processes, where what is to be controlled are process variables such as temperature, pressure, flow rate, and liquid level, in addition to product characteristics such as polymer viscosity, co-polymer composition, mole fraction of light material in an overhead distillation product, etc., the concept of “process control” takes on a somewhat different meaning. Let us begin by representing the variable to be controlled as:

$$y(k) = \eta(k) + e(k) \quad (22.50)$$

where  $\eta(k)$  is the true but unknown value of process variable;  $e(k)$  is noise, consisting of measurement error, usually random, and other unpredictable components;  $k = 1, 2, 3, \dots$  is a sampling time index.

The objective in chemical process control is to maintain the process variable as close as possible to its desired target value  $y_d(k)$ , in the face of possible



**FIGURE 22.15:** The EWMA chart for the original Mooney viscosity data using the same characteristics as in Fig 22.14. There are no non-conforming points.

systematic variations in the true value  $\eta(k)$ , and inherent random variations in the observed measurements  $y(k)$ . There are two fundamentally different perspectives of this problem, each leading naturally to a different approach philosophy:

1. The Statistical Process Control (SPC) perspective; and
2. The Engineering (or Automatic) Process Control (APC) perspective.

#### 22.4.2 Statistical Process Control (SPC) Perspective

The SPC techniques discussed previously take the following view of the control problem:  $\eta(k)$  tends to be constant, on target, with infrequent abrupt shifts in value; the shifts are attributable to “special causes” to be identified and eliminated; the noise term,  $e(k)$ , tends to behave like independently and identically distributed zero mean, gaussian random variables. Finally, taking control action is costly and should be done only when there is sufficient evidence in support of this need for action.

Taken to its logical conclusion, the implications of such a perspective is the following approach philosophy: observe each  $y(k)$  and analyze for infrequent “shifts;” take action only if a shift is detected with a pre-specified degree of confidence. It is therefore not surprising that the tools of SPC are embodied in the charts presented above: Shewhart, CUSUM, EWMA, etc. But it must be remembered that the original applications were in the discrete-parts manufacturing industry where the assumptions regarding the problem components as viewed from the SPC perspective are more likely to be valid.

### 22.4.3 Engineering/Automatic Process Control (APC) Perspective

The alternative APC perspective is that, left unattended,  $\eta(k)$  will wander and not remain on-target because of frequent, persistent, unavoidable and unmeasured/unmeasurable external disturbances that often arise from unknown sources; that regardless of underlying statistics, the contribution of the randomly varying noise term,  $e(k)$ , to the observation,  $y(k)$ , is minor compared to the contributions due to natural process dynamics and uncontrollable external disturbance effects. For example, the effect of outside temperature variations on the temperature profile in a refinery's distillation column that rises 130 ft into the open air will swamp any variations due to random thermocouple measurement errors. Finally, there is little or no cost associated with taking control action. For example, in response to an increase in the summer afternoon temperature, it is not costly to open a control valve to increase the cooling water flow rate to a distillation column's condenser; what is costly is *not* increasing the cooling and therefore allowing expensive light overhead material to be lost in the vent stream.

The natural implication of such a perspective is that every observed deviation of each  $y(k)$  from the desired  $y_d(k)$  is considered significant; as a result of which control action is implemented automatically at every sampling instant,  $k$ , according to a pre-designed control equation,

$$u(k) = f(\epsilon(k)) \quad (22.51)$$

where

$$\epsilon(k) = y_d(k) - y(k) \quad (22.52)$$

is the "feedback error" indicating the discrepancy between the observation  $y(k)$  and its desired target value,  $y_d(k)$ , and  $f(\cdot)$  is a control law that is based on specific design principles. For example, the standard (continuous time) Proportional-Integral-Derivative (PID) controllers operate according to

$$u(t) = K_c \left[ \epsilon(t) + \frac{1}{\tau_I} \int_0^t \epsilon(v) dv + \tau_D \frac{d\epsilon(t)}{dt} \right] \quad (22.53)$$

where  $K_c$ ,  $\tau_I$  and  $\tau_D$  are controller parameters chosen to achieve desired controller performance (see, for example, Ogunnaike and Ray, (1994) referenced earlier).

The engineering/automatic control philosophy therefore is to transfer variability from where it will "hurt" to where it will not — and as quickly as possible. This is to be understood in the sense that variability transmitted to the process variable,  $y(k)$ , (for example, the distillation column's temperature profile) if allowed to persist will adversely affect ("hurt") product quality; by adjusting the manipulated variable  $u(k)$  (for example, cooling water flow rate) in order to restore  $y(k)$  to its desired value, the variability that would have been observed in  $y(k)$  is thus transferred to  $u(k)$ , which usually is not a problem. (Variations in cooling water flow rate is typically not much of a concern

as for instance, variations in a distillation column's overhead product quality resulting from variations in the column temperature profile.)

Designing automatic controllers that are able to adjust manipulated variables  $u(k)$  to achieve engineering control objectives effectively is discussed in many textbooks on process control. On the surface, it appears as if SPC and APC approaches are diametrically opposed. However, a discussion in Chapter 28 of Ogunnaike and Ray (1994) puts the two approaches in perspective and shows how, given certain statistical models of chemical processes, the classical automatic controllers on the one hand, and the statistical process control charts on the other are in fact optimal stochastic controllers. For example:

1. For a “pure gain” process, for which

$$\eta(k) = Ku(k) \quad (22.54)$$

and with  $e(k) \sim N(0, \sigma^2)$ , the “minimum variance” control law is shown to be equivalent to the Shewhart charting paradigm;

2. If, instead of the zero mean gaussian noise model for  $e(k)$  above, the disturbance model is

$$d(k) = d(k - 1) + e(k) - \theta e(k - 1) \quad (22.55)$$

(known in time-series analysis as an integrated moving average, IMA(1,1) model), then the minimum variance controller is shown to be equivalent to using the EWMA chart to implement control action;

3. Finally, if the dynamic process model is first order, i.e.,

$$\eta(k) = a_1\eta(k - 1) + b_1u(k) + d(k) \quad (22.56)$$

or second order,

$$\eta(k) = a_1\eta(k - 1) + a_2\eta(k - 2) + b_1u(k) + b_2u(k - 1) + d(k) \quad (22.57)$$

where the disturbance model is as in Eq (22.55) above, then the minimum variance controller is shown to be exactly the discrete time version of the PID controller shown in Eq (22.53).

Additional details concerning these matters lie outside the intended scope of this chapter and the interested reader may consult the indicated reference. One aspect of that discussion that *will* be summarized here briefly concerns deciding when to choose one approach or the other.

#### 22.4.4 SPC or APC

The following three basic process and control attributes play central roles in choosing between the SPC approach or the APC approach:

1. *Sampling Interval:* This refers to how frequently the process output variable is measured; it is a process attribute best considered relative to the natural process response time. If a process with natural response time on the order of hours is sampled every minute, the dynamic characteristics of the process will be evident in the measurements and such measurements will be correlated in time. On the other hand, data sampled every hour from a process with a natural response time of minutes will most likely not show any dynamic characteristics and the observations are more likely to be uncorrelated.
2. *Noise (or disturbance) character:* This refers to the random or unexplainable part of the data. Is this due mainly to purely random variation with an occasional “special cause” shift, or is it due to systematic drifts and frequent “special cause” shifts?
3. *Cost of implementing control action:* Are control adjustments costly or mostly cost-free? This is the difference between shutting down a wafer polishing machine to adjust its settings versus opening and closing a control valve on a cold water supply line to a rector’s cooling jacket.

### When SPC is More Appropriate

The statistical process control approach is more appropriate when the three process and control attributes have the following characteristics:

1. *Sampling Interval:* When the sampling interval is large relative to the natural process response time, the assumption that the process mean is essentially constant and free of dynamics is reasonable. This may even allow the operator time to “find and eliminate” the assignable causes.
2. *Noise (or disturbance) character:* When the process is *not* subject to significant disturbances, the observed variability will then essentially be due to random events with infrequent “special cause” shifts.
3. *Cost of implementing control action:* When the cost of making control adjustments is significant, changes should be made only after there is conclusive evidence of a real need for the adjustment.

### When APC is More Appropriate

The automatic process control approach is more appropriate when the process and control attributes are characterized as follows:

1. *Sampling Interval:* When the sampling interval is small relative to the natural process response time, the natural process dynamics will be in evidence and the data will be serially correlated, contradicting a fundamental SPC assumption; uncorrected deviations from target will tend to persist; and there will be too many assignable causes, many of which cannot be corrected. (There is nothing, for example, that a refinery operator can do in the summer to “eliminate” the increasing outside temperature and prevent it from affecting the temperature profile in a distillation column that is exposed to the atmosphere.)
2. *Noise (or disturbance) character:* When the process is subject to persistent, significant disturbances, the observed variability will be due more to the effect of frequent “special cause” shifts than to purely random events that require no action.
3. *Cost of implementing control action:* When the cost of making control adjustments is negligible, control action should be taken according to well-designed control laws.

## 22.5 Process and Parameter Design

If acceptance sampling is taking action to rectify quality issues *after* manufacturing is complete, and process control is taking action *during* manufacturing, process and parameter design is concerned with taking action pre-emptively *before* manufacturing. According to this paradigm, wherever possible (and also economically feasible), operating parameters should be chosen to minimize the effects of uncontrollable factors that influence product quality variability. (If it is avoidable and the cost is not prohibitive, why expose a distillation column to the elements?)

### 22.5.1 Basic Principles

The primary ideas and concepts, due to Genichi Taguchi (born 1924), a Japanese engineer and statistician, involve using design of experiments to improve process operation *ahead of time*, by selecting those process parameters and operating conditions that are most conducive to robust process operation. With this paradigm, process variables are classified as follows:

1. *Response variables:* The variables of interest; quality indicators;

2. *Control factors:* The variables that affect the response; their levels can be decided by the experimenter, or process operator. (The reader familiar with chemical process control will recognize these variables as the manipulated variables);
3. *Noise factors:* The variables that affect the response but are not controllable. (Again, readers familiar with chemical process control will recognize these as the disturbance variables.)

Taguchi techniques are concerned with answering the question:

*At what level of control factors is the process response least susceptible to the effect of noise factors?*

This question is answered by conducting experiments using what are now known as Taguchi designs. These designs are structurally similar to the fractional factorial designs discussed in Chapter 19: they are orthogonal, and are all Resolution III designs. As in response surface designs, the primary objective is to find optimum settings for the control factors; but the optimization criteria are related not so much to the level of the response, but more importantly to the variability as well. The designs are therefore based on actively co-opting noise factors and using them to determine quantitatively, the most robust levels of control factors.

We are unable to accommodate any detailed discussions of these designs and methods within the scope of this single section; interested readers may consult the research monograph by Taguchi and Konishi, (1987)<sup>6</sup>, and/or the textbook by Ross<sup>7</sup>. What we are able to present is a brief theoretical rationale of the Taguchi concept of parameter design to show how it complements SPC and APC.

### 22.5.2 A Theoretical Rationale

Let  $Y$  represent the response of interest (for example, a product quality variable such as the Mooney viscosity of a commercial elastomer). The expected value and variance of this random variable are the usual  $\mu_Y$  and  $\sigma_Y^2$ , respectively. If  $y^*$  is the specified desired target value for  $Y$ , define

$$D = Y - y^* \quad (22.58)$$

---

<sup>6</sup>Taguchi, G. and Konishi, S., (1987), *Orthogonal Arrays and Linear Graphs*, Dearborn, MI, ASI Press.

<sup>7</sup>Ross, P.J. (1996). *Taguchi Techniques for Quality Engineering*, McGraw Hill, NY.

the deviation of this random variable from the target.  $D$  is itself a random variable with expected value,  $\delta$ , i.e.,

$$E(D) = \delta \quad (22.59)$$

Observe that by definition,

$$\delta = E(D) = E(Y) - y^* = \mu_Y - y^* \quad (22.60)$$

Let us now consider the following specific “loss function”:

$$L(y) = E(D^2) \quad (22.61)$$

as the loss incurred by  $Y$  deviating from the desired target (in this case, a squared error, or quadratic, loss function). By introducing Eq (22.58) into Eq (22.61), we obtain:

$$\begin{aligned} L(y) &= E[(Y - y^*)^2] \\ &= E\{[(Y - \mu_Y) + (\mu_Y - y^*)]^2\} \\ &= \sigma_Y^2 + \delta^2 \end{aligned} \quad (22.62)$$

This can be shown to represent an orthogonal decomposition of the squared deviation of the quality variable  $Y$ , from its target,  $y^*$ ; i.e.,

$$\begin{aligned} L(y) &= [\sigma_Y^2 + (\mu_Y - y^*)^2] \\ &\quad \text{Inherent Process + Process Operating} \\ &\quad \text{Variability} \qquad \qquad \qquad \text{Bias} \end{aligned} \quad (22.63)$$

Traditional SPC and/or engineering APC is concerned with, and can only deal with, the second term, by attempting to drive  $\mu_Y$  to  $y^*$  and hence eliminate the bias. Even if this can be achieved perfectly, the first term,  $\sigma_Y^2$ , still persists. Alternatively, this decomposition shows that the minimum achievable value for the loss function (achieved when  $\mu_Y = y^*$ ) is  $\sigma_Y^2$ . Thus even with a perfect control scheme,  $L(y) = \sigma_Y^2$ . Taguchi’s methodology is aimed at finding parameter settings to minimize this first term, by design. When the process is therefore operated at these optimum conditions, we can be sure that the minimum loss achieved with effective process control is the best possible. Without this, the manufacturer, even with the best control system, will incur product quality losses that cannot be compensated for any other way.

## 22.6 Summary and Conclusions

As with the previous chapter, this chapter has also been primarily concerned with showcasing one more application of probability and statistics that

has evolved into a full-scale subject matter in its own right. In this particular case, it is arguable whether the unparalleled productivity enjoyed by modern manufacturing will be possible without the tools of quality assurance and control, making the subject matter of this chapter one of the most influential applications of probability and statistics of the last century.

The presentation in three distinct units was a deliberate attempt to place in some historical perspective, the techniques that make up the core of quality assurance and control. Yet, fortuitously (or not) this demarcation also happens to coincide precisely with *where* along the manufacturing process time-line the technique in question is applicable. Thus, what we discussed first, acceptance sampling, with its “post-production” focus and applicability, is almost entirely a thing of the past (at least as a stand-alone quality assurance strategy). Our subsequent discussion of process control, the “during-production” strategy, covered the historical and modern incarnation of control charts, augmented with a brief summary of the automatic control paradigm—which included a discussion of how the two apparently opposite philosophies are really two perspectives of the same problem. In the final unit, we only provided the briefest of glances at the Taguchi techniques of parameter design, the “pre-production” strategy, choosing instead to emphasize the basic principles and rationale behind the techniques. Regardless of the successes of pre-production designs, however, process control will forever be an intrinsic part of manufacturing; the implementation ideas and techniques may advance, but the basic *concept* of monitoring process performance and making real-time, in-process adjustments to maintain control in the face of unavoidable, unpredictable, and potentially destabilizing variability, will always be a part of modern manufacturing.

We note, in closing, that because these quality control techniques arose from industrial needs, and were therefore developed exclusively for industrial manufacturing processes, they are so completely enmeshed with industrial practice that acquiring a *true* practical appreciation outside of the industrial environment is virtually impossible. To approximate the industrial experience of applying these techniques (especially to experience, first-hand, the real-time, sequential-in-time data structure that is intrinsic to these methods) we offer a few project assignments here in place of the usual exercises and applications problems.

## REVIEW QUESTIONS

1. To what objective is the subject matter of quality assurance and control devoted?
2. What characteristic of manufacturing processes makes quality assurance and control mandatory?
3. What are the three problems associated with assuring the quality of “mass produced” products? Which one did traditional quality assurance focus on?

4. What is the prevalent total quality management view of quality assurance?
5. What is acceptance sampling, and what is its defining characteristic?
6. What does acceptance sampling involve?
7. What is an “acceptance sampling by attribute” problem as opposed to an “acceptance sampling by variable” problem?
8. What is the Acceptable Quality Level (AQL)?
9. What is the Rejectable Quality Level (RQL), and by what alternative term is it also known?
10. What does it mean that a lot is of “indifferent quality”?
11. What is a “single sampling plan” as opposed to a double or multiple sampling plan?
12. What is the ideal sampling plan from a consumer’s perspective as opposed to what is ideal from a producer’s perspective?
13. Why is  $\alpha$ , the risk of committing a Type I error in hypothesis testing, also known as a producer’s risk?
14. Why is  $\beta$ , the risk of committing a Type II error in hypothesis testing, also known as a consumer’s risk?
15. What are sampling plans designed to do about the  $\alpha$  and  $\beta$  risks?
16. What is the operating characteristic (OC) curve and what is it mostly used for?
17. To generate a sampling plan, what four parameters must be specified?
18. What is the Binomial approximation OC curve and how is it different from the Poisson OC curve?
19. What is the shape of the ideal OC curve?
20. Why is acceptance sampling not considered a cost-effective quality assurance strategy?
21. What is the underlying philosophy behind process control as a quality assurance strategy? What primary issues are to be addressed in implementing such a strategy?
22. What is statistical process control (SPC)?
23. When is a process (or process variable) said to be in “statistical control”?

- 24.** What is “special cause” variability as opposed to “common cause” variability?
- 25.** In what way is SPC like hypothesis testing?
- 26.** What are control charts, and what three lines are found on all SPC charts?
- 27.** In traditional SPC what is the recommended corrective action for an “out-of-control” situation?
- 28.** What is the difference between control limits and specification limits?
- 29.** What is the Shewhart Xbar chart used for?
- 30.** What is the sampling distribution underlying the Shewhart chart?
- 31.** What are the characteristic components of the Shewhart chart?
- 32.** What is the S-chart used for and what are its main characteristics?
- 33.** What is the Xbar-R chart and how is it different from the I & MR chart?
- 34.** What is the P-chart used for and what are its main characteristics?
- 35.** What is the C-chart used for and what are its main characteristics?
- 36.** Why do basic SPC charts need enhancing?
- 37.** What are the Western Electric rules?
- 38.** What is the CUSUM chart and what are the principles behind it?
- 39.** What is the EWMA chart?
- 40.** What is the objective in chemical process control, and what are the two fundamentally different perspectives of the problem?
- 41.** What is the engineering/automatic process control perspective of the control problem?
- 42.** What is the engineering/automatic process control philosophy about transferring variability?
- 43.** What are the three basic process and control attributes to be used in choosing between statistical process control (SPC) and automatic process control (APC)?
- 44.** When is SPC the more appropriate approach to chemical process control?
- 45.** When is APC the more appropriate approach to chemical process control?

- 46.** If acceptance sampling is an “after-production” strategy, and process control is a “during-production” strategy, what sort of strategy is parameter design?
- 47.** In process and parameter design, what distinguishes “control” factors from “noise factors”?
- 48.** Taguchi techniques are concerned with answering what question?
- 49.** What are Taguchi designs?
- 50.** What is the squared error loss function?
- 51.** What are the two components of the squared error loss function and how are they connected to process control on one hand and parameter design on the other?

## PROJECT ASSIGNMENTS

### 1. Tracking the Dow

Even though many factors—some controllable, some not; some known, some unknown—contribute to the daily closing value of the Dow Jones Industrial Average (DJIA) index, it has been suggested that at a very basic level, the change in closing value from day to day in this index is distributed approximately as a zero mean random variable. Precisely what the distribution ought to be remains a matter of some debate.

Develop an SPC chart to track  $\delta(k) = \eta(k) - \eta(k-1)$ , where  $\eta(k)$  is the closing value of the Dow average on day  $k$ , with  $\eta(k-1)$  as the previous day’s value. From historical data during a typical period when the markets could be considered “stable,” determine base values for  $\mu_\delta$  and  $\sigma_\delta$ , or else assume that  $\mu_\delta = 0$  theoretically so that the chart will be used to identify any systemic departure from this postulated central value. Use the value estimated for  $\sigma_\delta$  (and a postulated probability model) to set the control limits objectively; track  $\delta(k)$  for 2 months with this chart. Should any point fall out of limits during this period, determine assignable causes where possible or postulate some. Present your results in a report.

Here are some points to consider about this project:

1. The I & MR and other similar charts are based on an implicit Gaussian distribution assumption. There have been arguments that financial variables such as  $\delta(k)$  are better modeled by the heavier-tailed Cauchy distribution (see Tanaka-Yamawaki (2003)<sup>8</sup> and Problem 18.16 in Chapter 18). Consider this in setting control limits and in using the limits to decide which deviations are to be considered as indicative of “out-of-control” fluctuation in the Dow average.
2. A possible alternative to consider is the EWMA chart which, as a moving average, is more likely to dampen excessive, but still “typical” fluctuations.
3. The DJIA is not the only index of the financial markets; in fact, many analysis

<sup>8</sup>Mieko Tanaka-Yamawaki, (2003). “Two-phase oscillatory patterns in a positive feedback agent model” *Physica A* 324, 380–387

argue that it is too narrow; that the Standard and Poors (S & P) 500 index provides a better gauge on the market. If time permits, consider a second chart simultaneously for the S & P 500 and assess whether or not the two indexes exhibit similar characteristics.

## **2. Diabetes and Process Control**

If you or a family member or a friend has diabetes, and the treatment procedure requires determining blood glucose level periodically, followed by self-administration of insulin as needed, consider a systematic process control approach to augment what is currently done.

Conduct a literature search on engineering control approaches for manual administration of insulin (see, e.g., Bequette and Desemone, (2004)<sup>9</sup>, and Zisser, *et al.*, (2005)<sup>10</sup>). Seek the assistance of a process control expert, if necessary. Determine a measurement protocol, appropriate control limits for the glucose level, and an appropriate control strategy. Implement this strategy over a period of 1 month. At the conclusion of the project, write a report on the planning, execution and the results. If possible, compare the performance of the systematic process control strategy with the strategy employed previously.

## **3. C-Chart for Sports Team**

The number of points or goals scored by a sports team in any particular game is a randomly varying quantity that depends on many largely uncontrollable factors. Use a *C*-chart to track the performance of your favorite team over a season. Determine base performance levels from historical data and use these to set the control limits. At the end of the season assess the team's performance from the perspective of the chart and write a brief report.

---

<sup>9</sup>Bequette, B. W. and J. Desemone, (2004). "Intelligent Dosing System": Need for Design and Analysis Based on Control Theory, *Diabetes Technology & Therapeutics*, 6(6): 868-873

<sup>10</sup>Zisser, H., Jovanovic L., Doyle, III F. J., Ospina P., Owens C. (2005), "Run-to-run control of mealrelated insulin dosing," *Diabetes Technol Ther* ; 7(1):48-57

# Chapter 23

## *Introduction to Multivariate Analysis*

|        |   |      |
|--------|---|------|
| 23.1   | Multivariate Probability Models .....                       | 978  |
| 23.1.1 | Introduction .....  | 978  |
| 23.1.2 | The Multivariate Normal Distribution .....                  | 979  |
| 23.1.3 | The Wishart Distribution .....                              | 980  |
| 23.1.4 | Hotelling's $T^2$ -Squared Distribution .....               | 982  |
| 23.1.5 | The Wilks Lambda Distribution .....                         | 982  |
| 23.1.6 | The Dirichlet Distribution .....                            | 983  |
| 23.2   | Multivariate Data Analysis .....                            | 984  |
| 23.3   | Principal Components Analysis .....                         | 985  |
| 23.3.1 | Basic Principles of PCA .....                               | 985  |
|        | Data Preconditioning .....                                  | 986  |
|        | Problem Statement .....                                     | 986  |
|        | Determining the Principal Components and Scores .....       | 987  |
| 23.3.2 | Main Characteristics of PCA .....                           | 989  |
|        | Some important results and implications .....               | 990  |
|        | Properties of PCA Transformation .....                      | 990  |
| 23.3.3 | Illustrative example .....                                  | 991  |
|        | Problem Statement and Data .....                            | 991  |
|        | PCA and Results .....                                       | 991  |
| 23.3.4 | Other Applications of PCA .....                             | 999  |
|        | Multivariate Process Monitoring .....                       | 999  |
|        | Model Building in Systems Biology .....                     | 1000 |
| 23.4   | Summary and Conclusions .....                               | 1002 |
|        | REVIEW QUESTIONS .....                                      | 1003 |
|        | PROJECT ASSIGNMENT .....                                    | 1004 |
|        | Principal Components Analysis of a Gene Expression Data Set | 1004 |

*The people are a many-headed beast*

Horace (65–8 BC)

To be sure, many practical problems involving randomly varying phenomena often manifest in the form of the single, isolated random variable,  $X$ , we have been studying thus far. Such single random variables—be they continuous like the yield obtainable from a chemical process, and the reliability of a microwave oven, or discrete like the number of flaws found on a manufactured sheet of glass—are mathematically characterized in terms of the probability models developed phenomenologically in Chapters 8 and 9, and via optimization in Chapter 10. After finalizing the discussion on probability models in those chapters, we have since focussed on how the models are used to solve problems of interest. However, it is also true that in a good number of practical problems, the phenomenon of interest involves multiple jointly distributed random variables, presenting a new set of challenges.

Conceptually, the principles that served us well with univariate problems remain unchanged: obtain a mathematical characterization (in the form of an appropriate probability model) and use it to solve the problem at hand. However, in the multivariate case, things are a bit more complicated. While the topic of multidimensional random variable characterization was broached in Chapter 5, beyond the quiet, unannounced appearance of the multinomial distribution in Chapter 8, not much has been said thus far about multivariate probability models and about intrinsically multivariate problems. That is about to change in this chapter. But multivariate analysis is a very rich and very broad topic; no single chapter can do it adequate justice. Therefore, the objective in this chapter is simply to alert the reader to the existence of these methods, and to demonstrate in a brief overview, how to generalize what has been presented in earlier chapters about probability models and statistical analysis to the multivariate case. We intend to present no more than a few important, hand-picked multivariate probability models (pdfs of vector-valued random variables), indicating their applications primarily by analogy to their univariate counterparts. (The reader will not be surprised to find that the role of vectors and matrices in facilitating the scaling up of scalar algebra to multidimensional linear algebra is reprised here, too.) We also introduce principal components analysis (PCA) as a representative methodology for multivariate exploratory data analysis, illustrating the technique and presenting some applications.

## 23.1 Multivariate Probability Models

### 23.1.1 Introduction

A multivariate probability model is the joint pdf of the multivariate random variable (or random vector)  $X = (X_1, X_2, \dots, X_n)$ . Conceptually, it is a direct extension of the single variable pdf,  $f(x)$ . As defined in Chapter 5, each component  $X_i$  of the random vector is itself a random variable with its own marginal pdf,  $f_i(x_i)$ ; and in the *special* case when these component random variables are independent, the joint pdf is obtained as a product of these individual marginal pdf's, i.e.,

$$f(\mathbf{x}) = \prod_{i=1}^n f_i(x_i) \quad (23.1)$$

as we saw in Chapters 13 and 14 while discussing sampling and estimation theory. In general, however, these constituent elements  $X_i$  are *not* independent, and the probability model will be more complex.

Here are some practical examples of multivariate random variables:

1. *Student SAT scores:* The SAT score for each student is a triplet of numbers:  $V$ , the score on the verbal portion,  $Q$ , the score on the quantitative portion, and  $W$ , the score on the writing portion. For a population of students, the score obtained by any particular student is therefore a three-dimensional random variable with components  $X_1$ ,  $X_2$  and  $X_3$  representing the individual scores on the verbal, quantitative and writing portions of the test respectively. Because students who do well in the verbal portion also tend to do just as well in the writing portion,  $X_1$  will be correlated with  $X_3$ . The total score,  $T = X_1 + X_2 + X_3$  is itself also a random variable.
2. *Product quality characterization of glass sheets:* Consider the case where the quality of manufactured glass sheets sold specifically into certain markets is characterized in terms of two quantities:  $X_1$ , representing the number of inclusions found on the sheet (see Chapter 1);  $X_2$ , representing “warp,” the extent to which the glass sheet is not flat (measured as an average curvature angle). This product quality variable is therefore a two-dimensional random variable. Note that one of the two components is discrete while the other is continuous.
3. *Market survey:* Consider a market evaluation of a several new products against their respective primary incumbent competitors: each subject participating in the market survey compares two corresponding products and gives a rating of 1 to indicate a preference for the new challenger, 0 if indifferent, and  $-1$  if the incumbent is preferred. The result of the market survey for each new product is the three-dimensional random variable with components  $X_1$ , the number of preferences for the new product,  $X_2$ , the number of “indifferents,” and  $X_3$ , the number of preferences for the incumbent.

The multinomial model is an example of a multivariate probability model; it was presented in Chapter 8, as a direct extension of the binomial probability model. We now present some other important multivariate probability models. A more complete catalog is available in Kotz *et al.* (2000)<sup>1</sup>

### 23.1.2 The Multivariate Normal Distribution

The joint pdf of the  $p$ -dimensional random variable,  $X = (X_1, X_2, \dots, X_p)^T$  for which each component random variable,  $X_i$ , is normally distributed, i.e.,  $X_i \sim N(\mu_i, \sigma_i^2)$ , is given by:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\} \quad (23.2)$$

---

<sup>1</sup>S. Kotz, N. Balakrishnan, and N. L. Johnson, (2000). *Continuous Multivariate Distributions*, Wiley, New York.

where  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$  are, respectively, the mean vector and the covariance matrix of the random vector,  $X$ , defined by:

$$\boldsymbol{\mu} = E(X) \quad (23.3)$$

$$\boldsymbol{\Sigma} = E[(X - \boldsymbol{\mu})(X - \boldsymbol{\mu})^T] \quad (23.4)$$

This is a direct extension of the Gaussian pdf given in Eq (9.125), with the vector  $\boldsymbol{\mu}$  taking the place of the single population mean,  $\mu$ , and  $\boldsymbol{\Sigma}$  taking the place of the population variance,  $\sigma^2$ . Thus, the important role of the Gaussian distribution in univariate analysis is played in multivariate analysis by the multivariate normal distribution. This particular  $k$ -dimensional vector of random variables is then said to possess a multivariate normal distribution, which we will represent as  $X \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ .

In the special case where  $p = 2$ , so that  $X = (X_1, X_2)^T$  with  $X_1 \sim N(\mu_1, \sigma_1^2)$  and  $X_2 \sim N(\mu_2, \sigma_2^2)$ , the result is the important bivariate normal distribution,  $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , for which

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad (23.5)$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} \quad (23.6)$$

Because of symmetry,  $\sigma_{12} = \sigma_{21}$ , and

$$\sigma_{12} = \rho\sigma_1\sigma_2 \quad (23.7)$$

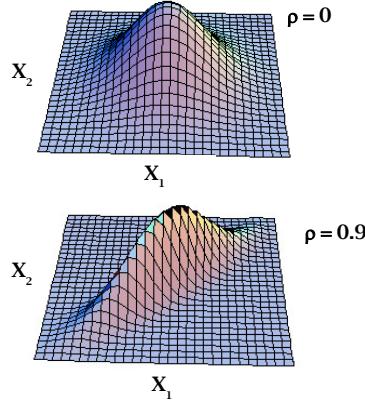
with  $\rho$  as the correlation coefficient. Eq (23.2) can be written out explicitly in this case to give:

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\{-U\} \quad (23.8)$$

with

$$U = \frac{1}{2(1-\rho)^2} \left[ \frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} - 2\rho \frac{(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} \right] \quad (23.9)$$

Fig 23.1 shows plots of the bivariate Gaussian distribution for  $\rho = 0$  (top) and  $\rho = 0.9$  (bottom), respectively. When  $\rho = 0$ , the two random variables are uncorrelated and the distribution is symmetric in all directions; when  $\rho = 0.9$ , the random variables are strongly positively correlated and the distribution is narrow and elongated along the diagonal.



**FIGURE 23.1:** Examples of the bivariate Gaussian distribution where the two random variables are uncorrelated ( $\rho = 0$ ) and strongly positively correlated ( $\rho = 0.9$ ).

### 23.1.3 The Wishart Distribution

Let  $\mathbf{X}$  be the  $n \times p$  matrix consisting of a random sample of size  $n$  drawn from a  $p$ -dimensional (i.e.,  $p$ -variate) normal distribution with zero mean and covariance matrix  $\Sigma$ , i.e.,  $\mathbf{X}_i^T \sim N_p(\mathbf{0}, \Sigma)$ ;  $i = 1, 2, \dots, n$ .

The elements of the  $p \times p$  dispersion matrix,  $\mathbf{V}$ , defined by:

$$\mathbf{V} = \mathbf{X}^T \mathbf{X} = \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^T \quad (23.10)$$

follow the Wishart distribution,  $W_p(\Sigma, n)$ , with  $n$  degrees of freedom, named for the Scottish statistician, John Wishart (1898-1956), who first developed the probability model. The pdf is:

$$f(\mathbf{V}) = \frac{|\mathbf{V}|^{(n-p-1)/2} \exp\{Tr(\mathbf{V}\Sigma^{-1})\}}{2^{np/2} \pi^{p(p-1)/4} |\Sigma|^{n/2} \prod_{i=1}^p \Gamma\left(\frac{n-i}{2}\right)} \quad (23.11)$$

where  $\Gamma(\cdot)$  is the Gamma function,  $||$  indicates the matrix determinant, and  $Tr()$  the trace of the matrix.

The Wishart distribution is a multivariate generalization of the  $\chi^2$  distribution. For the single variable case, where  $p = 1$  and  $\Sigma = 1$ , the Wishart distribution reduces to the  $\chi^2(n)$  distribution. The expected value of  $\mathbf{V}$  is:

$$E(\mathbf{V}) = n\Sigma \quad (23.12)$$

so that if  $\mathbf{S}$  is the sample covariance matrix for each of the  $p$ -dimensional normal random variables  $\mathbf{X}_i$ , then the distribution of  $\mathbf{S}$  is  $\frac{1}{n} W_p(\Sigma, n - 1)$ .

Therefore, the role played by the general  $\chi^2(r)$  distribution in univariate

statistical analysis is played by the Wishart distribution in multivariate analysis. Additional information about this distribution is available in Mardia, *et al.* (1979)<sup>2</sup>.

### 23.1.4 Hotelling's $T$ -Squared Distribution

Let the vector  $\mathbf{x}$  be a realization of a  $p$ -variate Gaussian random variable,  $MN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , and let  $\mathbf{S}$  be the sample covariance matrix obtained from  $n$  samples of the  $p$  elements of this vector; i.e.,

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T \quad (23.13)$$

where the vector average is defined as

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \quad (23.14)$$

Then from the previous section, we know that  $\mathbf{S} \sim \frac{1}{n} W_p(\boldsymbol{\Sigma}, n-1)$ . The statistic

$$T^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu})^T \mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \quad (23.15)$$

has the Hotelling's  $T$ -squared distribution  $T^2(p, n-1)$ . This distribution is a direct extension of the Student's  $t$ -distribution to the multivariate case; it is named for Harold Hotelling (1895–1973), the American mathematical statistician and econometrician who first developed the mathematical model<sup>3</sup>. Furthermore, it can be shown that:

$$\frac{n-p}{p(n-1)} T^2(p, n-1) \sim F(p, n-p) \quad (23.16)$$

where  $F(p, n-p)$  is the  $F$ -distribution with degrees of freedom  $p$  and  $n-p$ . Thus, the roles played by the Student's  $t$ -distribution and  $T$ -statistic in univariate analysis are played in multivariate analysis by the Hotelling's  $T^2$  distribution and statistic.

### 23.1.5 The Wilks Lambda Distribution

The multivariate generalization of the  $F$ -distribution is the Wilks Lambda distribution, named for the Princeton University mathematical statistician, Samuel S. Wilks (1906–1964). By direct analogy with its univariate counterpart, let the dispersion matrices  $\mathbf{U}$  and  $\mathbf{V}$  be independent and have the

---

<sup>2</sup>K. V. Mardia, Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis*, Academic Press, Duluth, London.

<sup>3</sup>H. Hotelling (1931). "The generalization of Student's ratio," *Ann. Math. Statist.* 2, 360378.

respective Wishart distributions,  $W_p(\mathbf{I}, n)$  and  $W_n(\mathbf{I}, m)$ ; i.e., each associated covariance matrix is the identity matrix,  $\mathbf{I}$ . Then the ratio  $\lambda$ , defined as:

$$\lambda = \frac{|\mathbf{U}|}{|\mathbf{U}| + |\mathbf{V}|} \quad (23.17)$$

has the Wilks  $\Lambda(p, n, m)$  distribution. It can be shown (see the Mardia *et al.*, reference (Reference 2)) that this distribution can be obtained as the distribution of a product of independent beta  $B(\alpha, \beta)$  random variables  $\xi_i$  where

$$\alpha = \frac{m+i-p}{2}; \beta = \frac{p}{2} \quad (23.18)$$

for  $m \geq p$ , i.e., if  $\xi_i \sim B(\alpha, \beta)$ , then

$$\prod_{i=1}^m \xi_i \sim \Lambda(p, n, m) \quad (23.19)$$

What the  $F$ -distribution is to the Student's  $t$ -distribution in univariate analysis, the Wilks  $\Lambda$  distribution is to Hotelling's  $T^2$  distribution.

### 23.1.6 The Dirichlet Distribution

The Dirichlet distribution, named for the German mathematician Johann Peter Gustav Lejeune Dirichlet (1805–1859), is the multivariate extension of the univariate Beta distribution. It arises as follows: Let  $Y_1, Y_2, \dots, Y_{k+1}$  be mutually independent random variables each with marginal  $\text{Gamma}(\alpha_1, 1)$  pdfs, i.e.,

$$f_i(y_i) = \frac{1}{\Gamma(\alpha_i)} y_i^{\alpha_i-1} e^{-y_i} \quad (23.20)$$

Define  $k$  ratios,

$$X_i = \frac{Y_i}{\sum_{j=1}^{k+1} Y_j}; \quad i = 1, 2, \dots, k \quad (23.21)$$

It can be shown that the joint pdf for the  $k$ -dimensional random variable  $X = (X_1, X_2, \dots, X_k)$  is given by:

$$f(\mathbf{x}) = \frac{\Gamma(\alpha_1 + \alpha_2 + \dots + \alpha_k)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\dots\Gamma(\alpha_k)} x_1^{\alpha_1-1} \dots x_{k-1}^{\alpha_{k-1}-1} (1 - x_2 - x_3 - \dots - x_{k-1})^{\alpha_k-1} \quad (23.22)$$

where, by definition,  $0 < x_i < 1$  and

$$x_1 + x_2 + \dots + x_k < 1 \quad (23.23)$$

Clearly, for  $k = 1$ , the pdf reduces to the beta distribution.

The most important characteristics of the distribution are as follows: The elements  $\mu_i$  of the mean vector,  $\boldsymbol{\mu}$ , are given by:

$$\mu_i = \frac{\alpha_i}{\sum_{i=1}^k \alpha_i} = \frac{\alpha_i}{\alpha} \quad (23.24)$$

The diagonal elements,  $\sigma_i^2$ , of the covariance matrix,  $\Sigma$ , are given by:

$$\sigma_i^2 = \frac{\alpha_i(\alpha - \alpha_i)}{\alpha^2(\alpha + 1)} \quad (23.25)$$

while the symmetric, off-diagonal elements are given by:

$$\sigma_{ij}^2 = \sigma_{ji}^2 = \frac{-\alpha_i\alpha_j}{\alpha^2(\alpha + 1)} \quad (23.26)$$

This distribution has found application, for example, in wildlife population studies in which  $n$  species (animals, plants, etc) reside in a given geographical area, the proportion of the area occupied by each species,  $\phi_1, \phi_2, \dots, \phi_n$  tend to follow a symmetric Dirichlet distribution. In engineering, the distribution has been used to model the activity times in a PERT (Program Evaluation and Review Technique) network. In particular, a Dirichlet distribution for the entire network can be used to derive an upper bound for a project's completion time<sup>4</sup>. It is also used as the conjugate prior distribution for Bayesian estimation of multinomial distribution parameters. The Kotz *et al.*, reference (Reference 1) contains additional details about applications of the Dirichlet distribution.

## 23.2 Multivariate Data Analysis

In what follows, unless stated otherwise, we will consider data consisting of  $n$  variables, ( $j = 1, 2, \dots, n$ ) and  $m$  samples of each variable, to give a data matrix  $\mathbf{X}$  that is of dimension  $m \times n$ , i.e.,

$$\begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_m \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix} \quad (23.27)$$

The variables can be correlated; furthermore, we impose no distributional assumptions on the random variables (at least for now).

In general, it is also possible to have another data block,  $\mathbf{Y}$ , (responses), consisting of  $n_y$  variables and  $m$  samples, where  $\mathbf{Y}$ , is related to  $\mathbf{X}$  (predictors).

Such data arise in various applications, from astrophysics to computer network traffic to chemical processes and molecular biology. The typical objectives of the analysis of such data include, but are not limited to:

---

<sup>4</sup>Monhor, D.,(1987). An approach to PERT: Application of Dirichlet distribution, *Optimization*, 18 113118.)

1. Identification, extraction and quantification of essential features in the data blocks;
2. Data Compression: Reducing data dimensionality “optimally” to a smaller number of essential (independent) components, and subsequent analysis from the perspective of the reduced data space;
3. Modeling: obtaining the “best possible” linear relationship between  $\mathbf{Y}$  (responses) and  $\mathbf{X}$  (predictors).

The typical applications in the chemical process industry include multivariate calibration and classification in analytical chemistry, process monitoring, data visualization and fault detection in manufacturing processes. Some applications in molecular biology include data-driven modeling and analysis of signal transduction networks; and characterization of the conformational space of proteins.

Of the various techniques of multivariate data analysis, including Multiple Linear Regression (MLR), Principal Component Regression (PCR), Partial Least Squares Regression (PLSR), we shall focus attention in the rest of the chapter only on Principal Components Analysis (PCA), since it provides most of the foundational elements of multivariate data analysis, from which the interested reader can then launch forays into the other aspects.

---

### 23.3 Principal Components Analysis

Principal components analysis (PCA), first proposed in 1901 by the British mathematical statistician, Karl Pearson (1857–1936)<sup>5</sup>, is a technique for transforming a data set consisting of a (large) number of possibly correlated variables from the original coordinate system into a new and usually more informative one. This new coordinate system consists of a smaller number of uncorrelated variables called *principal components*. The new set of variables forming the new coordinate system are not only mutually uncorrelated, they also represent a useful redistribution of the information in the data. This is because the new set of variables are ordered such that the greatest amount of variability is now captured along the first axis, the next greatest by the second, and on down to the last few axes with little or no information left to capture. PCA is therefore useful for data simplification (dimensionality reduction) because it allows the data to be described with only the first few truly informative component axes; it is also therefore useful for exposing hidden (latent) structures contained in the original data set.

---

<sup>5</sup>K. Pearson, (1901) “On Lines and Planes of Closest Fit to Systems of Points in Space.” *Phil. Magazine* 2 (6) 559–572. <http://stat.smmu.edu.cn/history/pearson1901.pdf>.

### 23.3.1 Basic Principles of PCA

#### Data Preconditioning

PCA is “scale-dependent,” with larger numerical values naturally accorded more importance, whether deserved or not. To eliminate any such undue influence (especially those arising from differences in the units in which different variables are measured), each data record can “mean-centered” and “scaled” (i.e., normalized) prior to carrying out PCA. But this is problem dependent.

Let the original data set consist of  $n$ -column vectors  $\mathbf{x}_1^\circ, \mathbf{x}_2^\circ, \dots, \mathbf{x}_n^\circ$ , each containing  $m$  samples, giving rise to the raw data matrix  $\mathbf{X}^\circ$ . If the mean and standard deviation for each column are  $\bar{x}_i, s_i$ , then each variable  $i = 1, 2, \dots, n$ , is normalized as follows:

$$\mathbf{x}_i = \frac{\mathbf{x}_i^\circ - \bar{x}_i}{s_i} \quad (23.28)$$

The resulting  $m \times n$  pre-treated data matrix,  $\mathbf{X}$ , consisting of columns of data each with zero mean and unit variance.

#### Problem Statement

We begin by contemplating the possibility of an orthogonal decomposition of  $\mathbf{X}$  by expanding in a set of  $n$ -dimensional orthonormal basis vectors,  $\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \dots, \mathbf{p}_n$ , according to:

$$\mathbf{X} = \mathbf{t}_1 \mathbf{p}_1^T + \mathbf{t}_2 \mathbf{p}_2^T + \dots + \mathbf{t}_n \mathbf{p}_n^T \quad (23.29)$$

in such a way that

$$\mathbf{p}_i^T \mathbf{p}_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \quad (23.30)$$

along with

$$\mathbf{t}_i^T \mathbf{t}_j = \begin{cases} \lambda_i & i = j \\ 0 & i \neq j \end{cases} \quad (23.31)$$

where  $\mathbf{t}_i$  ( $i = 1, 2, \dots, n$ ), like  $\mathbf{x}_i$ , are  $m$ -dimensional vectors.

In particular, if the  $k$ -term truncation of the complete  $n$ -term expansion,

$$\mathbf{X} = \mathbf{t}_1 \mathbf{p}_1^T + \mathbf{t}_2 \mathbf{p}_2^T + \dots + \mathbf{t}_k \mathbf{p}_k^T + \mathbf{E} \quad (23.32)$$

is such that the resulting residual matrix  $\mathbf{E}$ ,

$$\mathbf{E} = \sum_{i=k+1}^n \mathbf{t}_i \mathbf{p}_i^T \quad (23.33)$$

contains only random noise, such an expansion would have then provided a  $k$ -dimensional reduction of the data. It implies that  $k$  components are sufficient to capture all the useful variation contained in the data matrix.

The basis vectors,  $\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \dots, \mathbf{p}_k$ , are commonly referred to as “loading vectors;” they provide an alternative coordinate system for viewing the data.

The associated  $m$ -dimensional “weighting” vectors,  $\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3, \dots, \mathbf{t}_k$ , are the corresponding “score vectors” of the data matrix; they represent how the data will now appear in the principal component loading space, viewed from the perspective of the induced new coordinate system.

Together,  $\mathbf{t}_i$ ,  $\mathbf{p}_i$ , and  $k$  tell us much about the information contained in the data matrix  $\mathbf{X}$ . Principal components analysis of the data matrix,  $\mathbf{X}$ , therefore involves the determination of  $\mathbf{t}_i$ ,  $\mathbf{p}_i$ , and  $k$ .

### Determining the Principal Components and Scores

Unlike other orthogonal transforms (e.g., Fourier transforms) where the basis functions are known *a-priori* (sines and cosines for Fourier transforms), the basis vectors for PCA are not given ahead of time; they must be determined from the data itself. Thus, with PCA, *both* the basis function set as well as the “coefficients” of the transform are to be computed simultaneously. The following is one of several approaches for handling this problem; it is based on vector optimization.

We begin from

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \dots + \mathbf{t}_k\mathbf{p}_k^T + \mathbf{E} \quad (23.34)$$

and before engaging in any optimization, we define the  $n \times n$  symmetric matrix

$$\Phi_n = \mathbf{E}^T \mathbf{E} \quad (23.35)$$

and also the  $m \times m$  symmetric matrix

$$\Phi_m = \mathbf{E} \mathbf{E}^T \quad (23.36)$$

We now seek the  $\mathbf{t}_i$  and  $\mathbf{p}_i$  vectors that minimize the squared norm of the appropriate matrix  $\Phi_n$  or  $\Phi_m$ . Which matrix is appropriate depends on the dimensionality of the vector over which we are carrying out the minimization.

First, to determine each  $n$ -dimensional vector  $\mathbf{p}_i$ , since

$$\Phi_n = (\mathbf{X} - \mathbf{t}_1\mathbf{p}_1^T - \mathbf{t}_2\mathbf{p}_2^T - \dots - \mathbf{t}_k\mathbf{p}_k^T)^T (\mathbf{X} - \mathbf{t}_1\mathbf{p}_1^T - \mathbf{t}_2\mathbf{p}_2^T - \dots - \mathbf{t}_k\mathbf{p}_k^T) \quad (23.37)$$

we may differentiate with respect to the  $n$ -dimensional vector  $\mathbf{p}_i^T$  obtaining:

$$\frac{\partial \Phi_n}{\partial \mathbf{p}_i^T} = 2\mathbf{t}_i^T (\mathbf{X} - \mathbf{t}_1\mathbf{p}_1^T - \mathbf{t}_2\mathbf{p}_2^T - \dots - \mathbf{t}_k\mathbf{p}_k^T) \quad (23.38)$$

which, upon setting to the  $n$ -dimensional row vector of zeros,  $\mathbf{0}^T$ , and simplifying, yields

$$\frac{\partial \Phi_n}{\partial \mathbf{p}_i^T} = \mathbf{t}_i^T \mathbf{X} - \lambda_i \mathbf{p}_i^T = \mathbf{0}^T \quad (23.39)$$

where the simplification arises from the orthogonality requirements on  $\mathbf{t}_i$  (see Eq (23.31)). The solution is:

$$\mathbf{p}_i^T = \frac{\mathbf{t}_i^T \mathbf{X}}{\lambda_i} \quad (23.40)$$

Note that this result is true for all values of  $i$ , and is independent of  $k$ ; in other words, we would obtain precisely the same result regardless of the chosen truncation. This property is intrinsic to PCA and common with orthogonal decompositions; it is exploited in various numerical PCA algorithms. The real challenge at the moment is that Eq (23.40) requires knowledge of  $\mathbf{t}_i$ , which we currently do not have.

Next, to determine the  $m$ -dimensional vectors  $\mathbf{t}_i$ , we work with  $\Phi_m$ ,

$$\Phi_m = (\mathbf{X} - \mathbf{t}_1\mathbf{p}_1^T - \mathbf{t}_2\mathbf{p}_2^T - \dots - \mathbf{t}_k\mathbf{p}_k^T)(\mathbf{X} - \mathbf{t}_1\mathbf{p}_1^T - \mathbf{t}_2\mathbf{p}_2^T - \dots - \mathbf{t}_k\mathbf{p}_k^T)^T \quad (23.41)$$

and differentiate with respect to the  $m$ -dimensional vector  $\mathbf{t}_i$  to obtain:

$$\frac{\partial \Phi_m}{\partial \mathbf{t}_i} = 2(\mathbf{X} - \mathbf{t}_1\mathbf{p}_1^T - \mathbf{t}_2\mathbf{p}_2^T - \dots - \mathbf{t}_k\mathbf{p}_k^T)\mathbf{p}_i \quad (23.42)$$

Equating to the  $m$ -dimensional column vector of zeros,  $\mathbf{0}$ , and simplifying, the result is:

$$\frac{\partial \Phi_m}{\partial \mathbf{t}_i} = \mathbf{X}\mathbf{p}_i - \mathbf{t}_i\mathbf{p}_i^T\mathbf{p}_i = \mathbf{0} \quad (23.43)$$

or,

$$\mathbf{t}_i = \frac{\mathbf{X}\mathbf{p}_i}{\mathbf{p}_i^T\mathbf{p}_i} = \mathbf{X}\mathbf{p}_i \quad (23.44)$$

where the simplification arises again from the orthogonality requirements on  $\mathbf{p}_i$  (see Eq (23.30)).

We now have two self-referential expressions: one for determining  $\mathbf{p}_i$  if  $\mathbf{t}_i$  is known, Eq (23.40); the other for determining  $\mathbf{t}_i$  if  $\mathbf{p}_i$  is known, Eq (23.44). But neither is currently known. To resolve this conundrum, we start from Eq (23.40) and substitute Eq (23.44) for  $\mathbf{t}_i$  to eliminate it and obtain:

$$\mathbf{p}_i^T = \frac{\mathbf{p}_i^T \mathbf{X}^T \mathbf{X}}{\lambda_i} \quad (23.45)$$

and if we let  $\mathbf{R} = \mathbf{X}^T \mathbf{X}$ , then Eq (23.45) simplifies to

$$\begin{aligned} \mathbf{p}_i^T \mathbf{R} &= \lambda_i \mathbf{p}_i^T \\ \text{or } \mathbf{p}_i^T (\mathbf{R} - \lambda_i \mathbf{I}) &= \mathbf{0} \end{aligned} \quad (23.46)$$

and finally, because both  $\mathbf{R}$  and  $\mathbf{I}$  are symmetric, we have

$$(\mathbf{R} - \lambda_i \mathbf{I})\mathbf{p}_i = \mathbf{0} \quad (23.47)$$

This equation is immediately recognized as the eigenvalue-eigenvector equation, with the following implications:

*The loading vectors  $\mathbf{p}_i$  are the eigenvectors of the matrix  $\mathbf{R} = \mathbf{X}^T \mathbf{X}$ , with the eigenvalues given by  $\lambda_i = \mathbf{t}_i^T \mathbf{t}_i$  (see eqn (23.31)).*

Thus, to carry out PCA:

1. Optionally mean-center and normalize the original data matrix  $\mathbf{X}^o$  to obtain the data matrix  $\mathbf{X}$ ;
2. Obtain the matrix  $\mathbf{R} = \mathbf{X}^T \mathbf{X}$ ; (if the data is mean-centered and normalized, this matrix is related to the correlation matrix; if the data is mean-centered only, then  $\mathbf{R}$  is related to the covariance matrix  $\Sigma$ , which is  $\frac{1}{m-1}\mathbf{R}$ );
3. Obtain the eigenvalues and corresponding eigenvectors of  $\mathbf{R}$ ; arrange the eigenvalues in descending order such that  $\lambda_1 > \lambda_2 > \dots > \lambda_n$ ; the corresponding eigenvectors are the loading vectors  $\mathbf{p}_i; i = 1, 2, \dots, n$ ;
4. Obtain the corresponding scores from Eq (23.44) by projecting the data matrix onto the loading vector  $\mathbf{p}_i$ .

Even though determining the truncation  $k$  is somewhat subjective, some methods for choosing this parameter are available, for example, in Malinowski (1991)<sup>6</sup> and Kritchman and Nadler (2008)<sup>7</sup>. For a chosen truncation,  $k < n$  from Eq (23.44), obtain:

$$[\mathbf{t}_1 \mathbf{t}_2 \dots \mathbf{t}_k] = \mathbf{X} [\mathbf{p}_1 \mathbf{p}_2 \dots \mathbf{p}_k] \quad (23.48)$$

or, simply

$$\mathbf{T} = \mathbf{X}\mathbf{P} \quad (23.49)$$

as the “principal component” transform of the data matrix  $\mathbf{X}$ . The  $k$  transformed variables  $\mathbf{T}$  are called the principal components scores.

The corresponding “inverse transform” is obtained from (23.29) as:

$$\hat{\mathbf{X}} = \mathbf{T}\mathbf{P}^T \quad (23.50)$$

with  $\hat{\mathbf{X}} = \mathbf{X}$  only if  $k = n$ ; otherwise  $\hat{\mathbf{X}}$  is a “cleaned up,” lower-rank version of  $\mathbf{X}$ . The difference,

$$\mathbf{E} = \mathbf{X} - \hat{\mathbf{X}} \quad (23.51)$$

is the residual matrix; it represents the residual information contained in the portion of the original data associated with the  $(n - k)$  loading vectors that were excluded from the transformation.

---

<sup>6</sup>Malinowski, E. R. (1991). *Factor Analysis in Chemistry*, John Wiley & Sons

<sup>7</sup>Kritchman, S. and B. Nadler (2008). “Determining the number of components in a factor model from limited noisy data.” *Chemometrics and Intelligent Laboratory Systems* 94(1): 19-32.

### 23.3.2 Main Characteristics of PCA

#### Some important results and implications

The determination of  $\mathbf{P}$  is an eigenvalue-eigenvector problem; the numerical computation is therefore typically carried out via singular value decomposition. The following expressions hold for  $\mathbf{P}$ :

$$\mathbf{P}^T \mathbf{R} \mathbf{P} = \mathbf{\Lambda} \quad (23.52)$$

$$\text{or } \mathbf{R} = \mathbf{P} \mathbf{\Lambda} \mathbf{P}^T \quad (23.53)$$

with the following implications:

1.  $\text{Tr}(\mathbf{R})$ , the trace of the matrix  $\mathbf{R}$ , the sum of the diagonal elements of the matrix  $\mathbf{R}$ , is equal to  $\text{Tr}(\mathbf{\Lambda})$ , the sum of the eigenvalues; i.e.,

$$\text{Tr}(\mathbf{R}) = \text{Tr}(\mathbf{\Lambda}) = \sum_{i=1}^n \lambda_i \quad (23.54)$$

2.  $\text{Tr}(\mathbf{R})$  is a measure of the total variance in the data block  $\mathbf{X}$ . From Eq (23.54), this implies that the sum of the eigenvalues is also a measure of the total variance in the data block. The fractional contribution of the  $j^{th}$  principal component to the overall variance in the data is therefore given by  $\lambda_j / (\sum_{i=1}^n \lambda_i)$ .

3. Similarly,

$$|\mathbf{R}| = \prod_{i=1}^n \lambda_i \quad (23.55)$$

If, therefore, the matrix  $\mathbf{X}^T \mathbf{X} = \mathbf{R}$  is of rank  $r < n$ , then only  $r$  eigenvalues are non-zero; the rest are zero, and the determinant will be zero. The matrix will therefore be singular (and hence non-invertible).

4. When an eigenvalue is not precisely zero, just small, the cumulative contribution of its corresponding principal component to the overall variation in the data will likewise be small. Such component may then be ignored. Thus, by defining the cumulative contribution of the  $j^{th}$  principal component as:

$$\Gamma_j = \frac{\sum_{i=1}^j \lambda_i}{\sum_{i=1}^n \lambda_i} \quad (23.56)$$

one may choose  $k$  such that  $\Gamma_{k+1}$  “does not add much” beyond  $\Gamma_k$ . Typically a plot of  $\Gamma_j$  versus  $j$ , known as a Scree plot, shows a “knee” at or after the point  $j = k$  (see Fig 23.3).

### Properties of PCA Transformation

As a technique for transforming multivariate data from one set of coordinates into a new, more informative set, here are some of the key properties of PCA:

1. Each loading vector  $\mathbf{p}_i^T$  is seen from Eq (23.40) to be a linear combination of the columns of the data matrix.
2. Each score,  $\mathbf{t}_i$ , is seen from Eq (23.44) as the projection of the data onto the basis vector  $\mathbf{p}_i$ . By choosing  $k < n$ , the data matrix  $\mathbf{X}$  is projected down to a lower dimensional space using  $\mathbf{p}_i^T$  as the basis for the projection.
3. For all intents and purposes, PCA “replaces” the original  $m \times n$  data matrix  $\mathbf{X}$  with a better conditioned  $m \times k$  matrix  $\mathbf{T}$  in a different set of coordinates in a lower dimensional sub-space of the original data space. The data may be “recovered” in terms of the original coordinates after eliminating the extraneous components from Eq (23.50), where  $\hat{\mathbf{X}}$  is now an  $m \times k$  matrix.
4. Any collinearity problem in  $\mathbf{X}$  is solved by this transformation because the resulting matrix  $\mathbf{T}$  is made up of orthogonal vectors so that  $\mathbf{T}^T\mathbf{T}$  not only exists, it is diagonal.

The principles and results of PCA are now illustrated with the following example.

#### 23.3.3 Illustrative example

Even to veterans of multivariate data analysis, the intrinsic linear combinations of variables can make PCA and its results somewhat challenging to interpret. This is in addition to the usual plotting and visualization challenge arising from the inherent multidimensional character of such data analysis. The problem discussed here has been chosen therefore specifically to demonstrate what principal components, scores and loadings mean in real applications, but in a manner somewhat more transparent to interpret<sup>8</sup>.

#### Problem Statement and Data

The problem involves 100 samples obtained from 16 variables,  $Y_1, Y_2, \dots, Y_{16}$ , to form a  $100 \times 16$  raw data matrix, a plot of which is shown in Fig 23.2. The primary objective is to analyze the data to see if the dimensionality could be reduced from 16 to a more manageable number; and to see if there are any patterns to be extracted from the data. Before going on, the reader is encouraged to take some time and examine the data plots for any visual clues regarding the characteristics of the complete data set.

---

<sup>8</sup>The problem has been adapted from an example communicated to me by M. J. Piovoso.

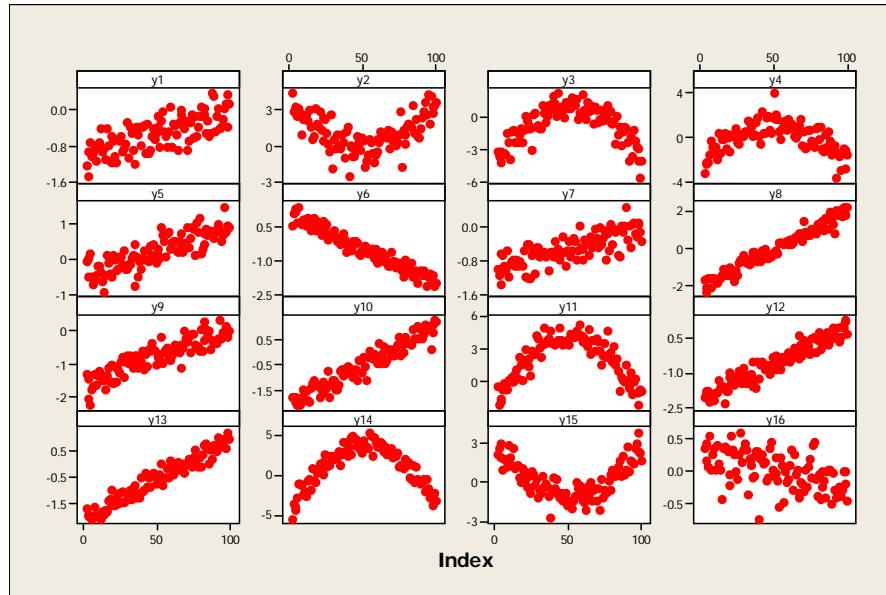


FIGURE 23.2: Plot of the 16 variables in the illustrative example data set.

### PCA and Results

The computations involved in PCA are routinely done with computer software; specifically with MINITAB, the sequence, **Stat > Multivariate > Principal Components >** opens a dialog box where the problem characteristics are specified. In particular, the data columns are selected (all 16 of them), along with the number of principal components to be computed (we select 6 to start with). The type of data matrix form to use is selected as “Correlation” (this is the scaled, mean-centered form; the alternative, “Covariance,” is mean-centered only). We chose to store the scores and the loadings (Eigenvalues). MINITAB also has several plotting options. The MINITAB results are shown below, first the eigenvalues and their respective contributions:

#### Eigenanalysis of the Correlation Matrix

| Eigenvalue | 7.7137 | 4.8076 | 0.7727 | 0.5146 |
|------------|--------|--------|--------|--------|
| Proportion | 0.482  | 0.300  | 0.048  | 0.032  |
| Cumulative | 0.482  | 0.783  | 0.831  | 0.863  |
| <hr/>      |        |        |        |        |
| Eigenvalue | 0.4517 | 0.3407 | 0.2968 | 0.2596 |
| Proportion | 0.028  | 0.021  | 0.019  | 0.016  |
| Cumulative | 0.891  | 0.913  | 0.931  | 0.947  |

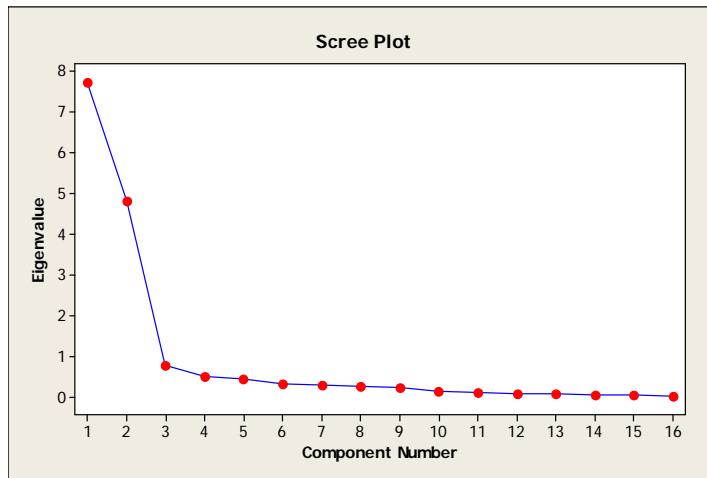
|            |        |        |        |        |
|------------|--------|--------|--------|--------|
| Eigenvalue | 0.2419 | 0.1608 | 0.1248 | 0.1021 |
| Proportion | 0.015  | 0.010  | 0.008  | 0.006  |
| Cumulative | 0.962  | 0.973  | 0.980  | 0.987  |
| <hr/>      |        |        |        |        |
| Eigenvalue | 0.0845 | 0.0506 | 0.0437 | 0.0342 |
| Proportion | 0.005  | 0.003  | 0.003  | 0.002  |
| Cumulative | 0.992  | 0.995  | 0.998  | 1.000  |

The principal components (the loading vectors) are obtained as follows:

| Variable | PC1    | PC2    | PC3    | PC4    | PC5    | PC6    |
|----------|--------|--------|--------|--------|--------|--------|
| y1       | 0.273  | 0.005  | -0.015 | 0.760  | 0.442  | 0.230  |
| y2       | 0.002  | -0.384 | 0.151  | -0.253 | 0.396  | -0.089 |
| y3       | 0.028  | 0.411  | 0.006  | -0.111 | 0.125  | 0.151  |
| y4       | -0.016 | 0.381  | 0.155  | 0.239  | 0.089  | -0.833 |
| y5       | 0.306  | -0.031 | -0.180 | -0.107 | 0.407  | 0.077  |
| y6       | -0.348 | 0.001  | 0.055  | 0.119  | 0.009  | 0.172  |
| y7       | 0.285  | 0.002  | -0.247 | 0.312  | -0.631 | 0.054  |
| y8       | 0.351  | -0.018 | -0.030 | -0.029 | 0.019  | 0.007  |
| y9       | 0.326  | 0.027  | -0.014 | -0.252 | -0.100 | -0.025 |
| y10      | 0.346  | -0.020 | -0.066 | -0.063 | -0.028 | -0.037 |
| y11      | 0.019  | 0.427  | -0.002 | -0.098 | 0.042  | 0.238  |
| y12      | 0.344  | -0.025 | 0.000  | -0.117 | 0.029  | -0.198 |
| y13      | 0.347  | -0.014 | 0.011  | -0.193 | -0.027 | -0.047 |
| y14      | 0.036  | 0.427  | 0.066  | -0.116 | 0.124  | 0.070  |
| y15      | -0.007 | -0.412 | 0.023  | 0.142  | -0.014 | -0.219 |
| y16      | -0.199 | 0.030  | -0.920 | -0.068 | 0.175  | -0.175 |

These results are best appreciated graphically. First, Fig 23.3 is a *Scree plot*, a straightforward plot of the eigenvalues in descending order. The primary characteristic of this plot is that it shows graphically how many eigenvalues (and hence principal components) are necessary to capture most of the variability in the data. This particular plot shows that after the first two components, not much else is important. The actual numbers in the eigenanalysis table show that almost 80% of the variability in the data is captured by the first two principal components; the third principal component contributes less than 5% following the 30% contribution from the second principal component. This is reflected in the very sharp “knee” at the point  $k + 1 = 3$  in the Scree plot. The implication therefore is that the information contained in the 16 variables can be represented quite well using two principal components, PC1 and PC2, shown in the MINTAB output table.

If we now focus on the first two principal components and their associated scores and loadings, the first order of business is to plot these to see what insight they can offer into the data. The individual scores and loading plots are particularly revealing for this particular problem. Fig 23.4 shows such a plot for the first principal component. It is important to remember that the scores indicate what the new data will look like in the transformed coordinates; in this case, the top panel indicates that in the direction of the first principal



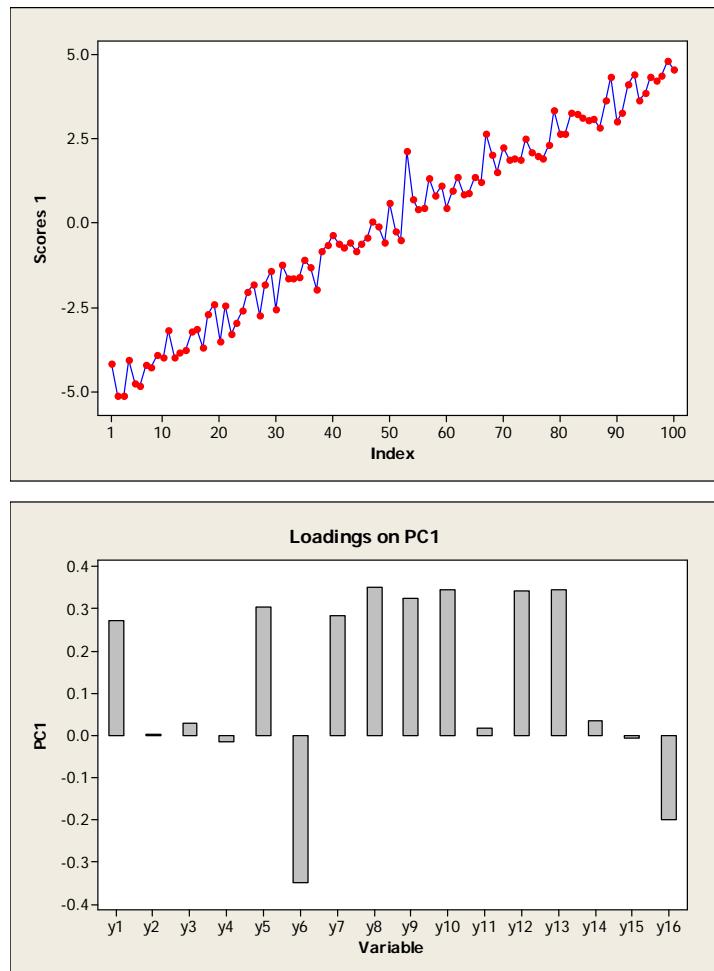
**FIGURE 23.3:** Scree plot showing that the first two components are the most important.

component, the data set is essentially linear with a positive slope. The loading plot indicates in what manner this component is represented in (or contributes to) each of the original 16 variables.

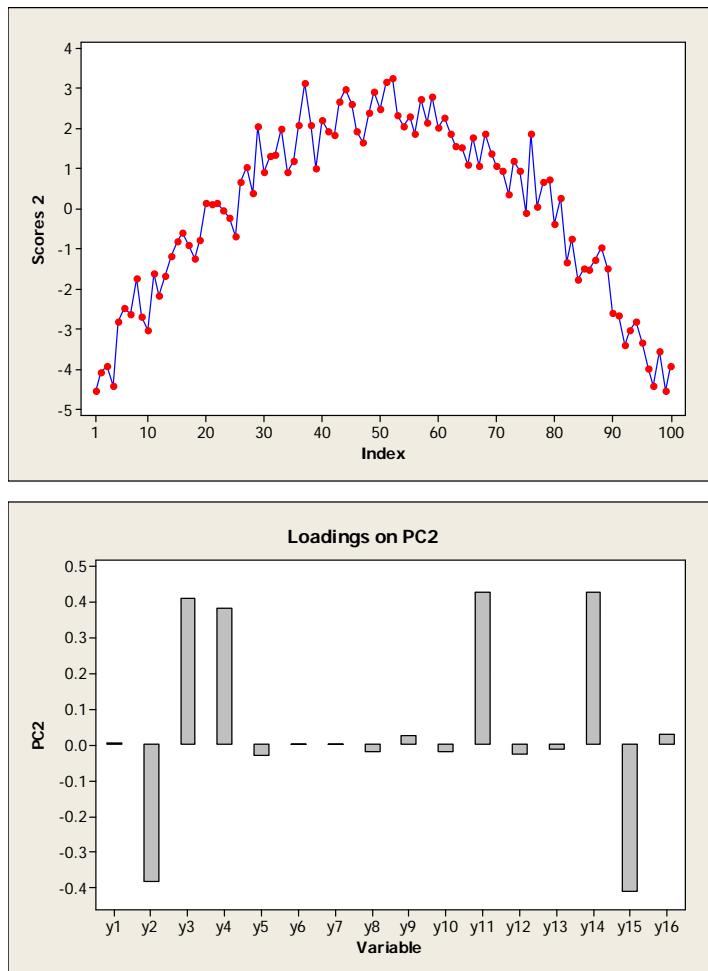
The corresponding plot for the second principal component is shown in Fig 23.5 where we observe another interesting characteristic: the top panel (the scores) indicates that in the direction of the second principal component, the data is a downward pointing quadratic; the bottom panel, the loadings associated with each variable, indicates how this quadratic component contributes to each variable.

Taken together, these plots indicate that the data consists of only two primary modes: PC1 is linear, and the more dominant of the two; the other, PC2, is quadratic. Furthermore, the loadings associated with PC1 indicate that the linear mode manifests negatively in two variables,  $Y_6$  and  $Y_{16}$ ; the variables for which the loadings are strong and positive (i.e.,  $Y_1, Y_5, Y_7, Y_8, Y_9, Y_{10}, Y_{12}$  and  $Y_{13}$ ) contain significant amounts of the linear mode at roughly the same level. For the other remaining 6 variables, the indication of Fig 23.4 is that they do not contain any of the linear mode. The story for PC2 is similar but complementary: the quadratic mode manifests negatively in two variables  $Y_2$  and  $Y_{15}$ , and positively in four variables,  $Y_3, Y_4, Y_{11}$  and  $Y_{14}$ . The quadratic mode contributes virtually nothing to the other variables.

It is now interesting to return to the original data set in Fig 23.2 to compare the raw data with the PCA results. It is now obvious that other than noise, the data sets consists of linear and quadratic trends only, some positive, some negative. The principal components reflect these precisely. For example, 10 of the 16 variables show the linear trends; the remaining 6 show the quadratic trend. The first principal component, PC1, reflects the linear trend as the



**FIGURE 23.4:** Plot of the scores and loading for the first principal component. The distinct trend indicated in the scores should be interpreted along with the loadings by comparison to the full original data set in Fig 23.2.



**FIGURE 23.5:** Plot of the scores and loading for the second principal component. The distinct trend indicated in the scores should be interpreted along with the loadings by comparison to the full original data set in Fig 23.2.

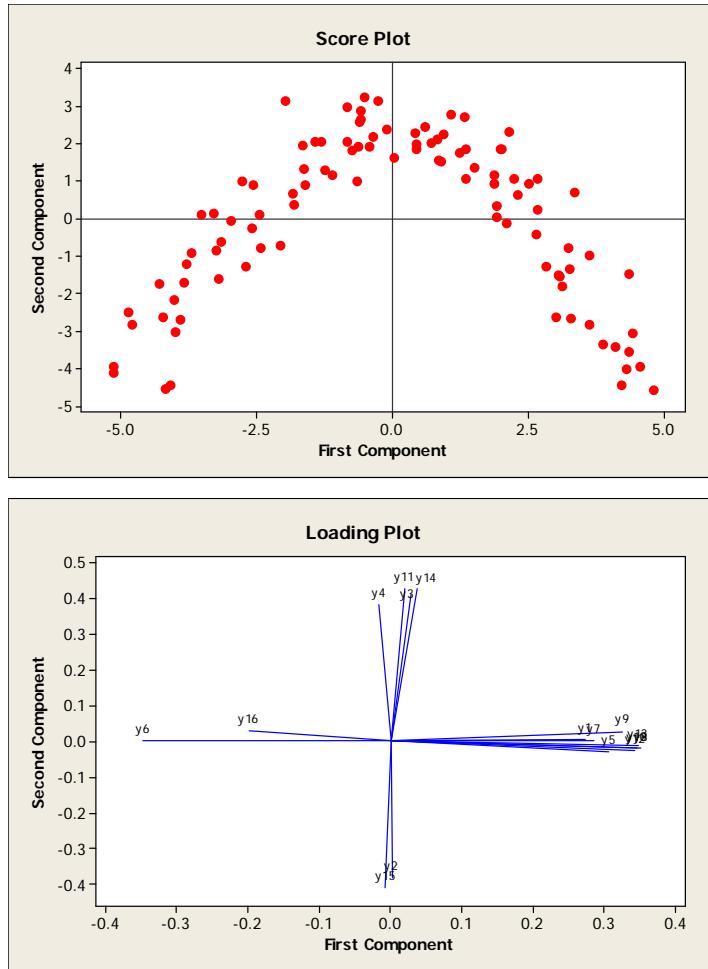
more dominant and the variables with the linear trends are all identified in the loadings of the PC1; no variable showing a quadratic trend is included in this set of loadings. Furthermore, among the variables showing the linear trend, the slope is negative in  $Y_6$  and in  $Y_{16}$ ; it is positive in the others. This is reflected perfectly in the loadings for PC1 where the values associated with the variables  $Y_6$  and  $Y_{16}$  are negative, but positive for the others. In the same manner the component capturing the downward pointing quadratic trend, PC2, is associated with two groups of variables: (i) the variables  $Y_2$  and  $Y_{15}$ , whose raw observations show an upward pointing quadratic (hence the negative values of the loadings associated with PC2); and (ii) the variables,  $Y_3, Y_4, Y_{11}$  and  $Y_{14}$ , which all show downward pointing quadratic trends; these all have positive values in the PC2 loadings.

Finally, we show in Fig 23.6, the two-dimensional score and loading plots for the first component versus the second. Such plots are standard fare in PCA. They are designed to show any relationship that might exist between the scores in the new set of coordinates, and also how the loading vectors of the first two principal components are related. For this specific example, the 2-D score plot indicates a distinct quadratic relationship between  $t_1$  and  $t_2$ . To appreciate the information encoded in this plot, observe that the scores associated with PC1 (shown in Fig 23.4) appear linear in the form  $t_1 = a_1 \mathbf{x}$  where  $\mathbf{x}$  represents the independent variable (because the data matrix has been mean-centered, there is no need for an intercept). Likewise, the second set of scores appear quadratic, in the form  $t_2 = a_2 \mathbf{x}^2$  (where the exponent is to be understood as a term-by-term squaring of the elements in the vector  $\mathbf{x}$ ) so that indeed  $t_2 = b t_1^2$  where  $b = a_2/a_1^2$ . This last expression, the relationship between the two scores, is what the top panel of Fig 23.6 is encoding.

The 2-D loading plot reveals any relationships that might exist between the new set of basis vectors constituting PC1 and PC2; it invariably leads to clustering of the original variables according to patterns in the data. In this particular case, this plot shows several things simultaneously: first, its North-South/West-East alignment indicates that in terms of the original data, these two principal components are pure components: the linear component in PC1 is pure, with no quadratic component; similarly, PC2 contains a purely quadratic component. The plot also indicates that the variables  $Y_6$  and  $Y_{16}$ , cluster together, lying on the negative end of PC1;  $Y_1, Y_5, Y_7, Y_8, Y_9, Y_{10}, Y_{12}$  and  $Y_{13}$  also cluster together at the positive extreme of the pure component PC1. The reader should now be able to interpret the vertical segregation and clustering of the variables showing the quadratic trends.

To summarize, PCA has provided the following insight into this data set:

- It contains only two modes: linear (the more dominant) and quadratic;
- The 16 variables each contain these modes in pure form: the ones showing the linear trend do not show the quadratic trend, and vice versa;
- The variables can be grouped into four distinct categories: (i) Negative linear ( $Y_6, Y_{16}$ ); (ii) Positive linear ( $Y_1, Y_5, Y_7, Y_8, Y_9, Y_{10}, Y_{12}$  and



**FIGURE 23.6:** Scores and loading plots for the first two components. Top panel: Scores plot indicates a quadratic relationship between the two scores  $t_1$  and  $t_2$ ; Bottom panel: Loading vector plot indicates that in the new set of coordinates, the original variables contain mostly pure components PC1 and PC2 indicated by a distinctive North/South and West/East alignment of the data vectors, with like variables clustered together according to the nature of the component contributions. Compare to the full original data set in Fig 23.2.

$Y_{13}$ ); (iii) Negative quadratic ( $Y_2$  and  $Y_{15}$ ); and (iv) Positive quadratic ( $Y_3, Y_4, Y_{11}$  and  $Y_{14}$ ).

It is of course rare to find problems for which the principal components are as pure as in this example. Keep in mind, however, that this example is a deliberate attempt to give the reader an opportunity to see first a transparent case where the PCA results can be easily understood. Once grasped, such understanding is then easier to translate to less transparent cases. For example, had one of the variables contained a mixture of the linear and quadratic trends, the extent of the mixture would have been reflected in the loadings for each of the scores: the length of the bar in Fig 23.4 would have indicated how much of the linear trend it contains, with the corresponding bar in Fig 23.5 indicating the corresponding relative amount of the quadratic trend. The 2-D loading vector for this variable will then lie at an angle (no longer horizontal or vertical) indicative of the relative contribution of the linear PC1 and that of PC2 to the raw data.

Additional information especially about implementing PCA in practice may be found in Esbensen (2002)<sup>9</sup>, Naes *et al.*, (2002)<sup>10</sup>, Brereton (2003)<sup>11</sup>, and in Massart *et al.*, (1988)<sup>12</sup>.

### 23.3.4 Other Applications of PCA

#### Multivariate Process Monitoring

When the task of ascertaining that a process is “in control” requires simultaneous monitoring of several process variables, the single variable charts discussed in the previous chapter will no longer work because of potential correlation between variables. PCA can be, and has been, used to handle this problem. Carrying out PCA on a “training” data set,  $\mathbf{X}$ , produces loading vectors,  $\mathbf{p}_i$ , the matrix of eigenvalues,  $\Lambda$ , and the sample covariance matrix,  $\mathbf{S}$  for typical operation. This exercise achieves two important objectives: (i) it takes care of any correlations in the variables by reordering the data matrix in terms of scores and orthogonal PC loadings; but more importantly, (ii) it provides a data-based model of what is considered normal operation. Each new sample can then be evaluated against “normal operation” by projecting unto the loading vectors and analyzing the result in the score space.

Thus, the new data samples,  $\tilde{\mathbf{X}}$ , are projected to obtain the corresponding scores  $\tilde{\mathbf{t}}_i$  and a new error matrix,  $\tilde{\mathbf{E}}$ , using the same loading vectors,  $\mathbf{p}_i$ ,

<sup>9</sup>K. H. Esbensen, (2002). *Multivariate Data Analysis-In practice* (5th Edition), Camo Process AS.

<sup>10</sup>Naes, T., T. Isaksson, T. Fearn and T. Davies (2002). *A user-friendly guide to Multivariate Calibration and Classification*. Chichester, UK, NIR Publications.

<sup>11</sup>Brereton, R. G. (2003). *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*, Wiley & Sons.

<sup>12</sup>Massart, D. L., B. G. M. Vandeginste, S. N. Deming, Y. Michotte and L. Kaufman (1988). *Chemometrics: A Textbook*. Amsterdam, Netherlands, Elsevier.

according to:

$$\tilde{\mathbf{X}} = \tilde{\mathbf{t}}_1 \mathbf{p}_1^T + \tilde{\mathbf{t}}_2 \mathbf{p}_2^T + \dots + \tilde{\mathbf{t}}_k \mathbf{p}_k^T + \tilde{\mathbf{E}} \quad (23.57)$$

Two statistics are used to assess the new data set against normal operation: the  $Q$  statistic, from the  $i^{th}$  row of  $\tilde{\mathbf{E}}$ ,

$$Q_i = \tilde{\mathbf{e}}_i^T \tilde{\mathbf{e}}_i \quad (23.58)$$

is the error sum of squares for the  $i^{th}$  sample (also known as the lack-of-model-fit statistic); it provides a measure of how well the  $i^{th}$  sample conforms to the PCA model and represents the distance between the sample and its projection onto the  $k$ -dimensional principal components space. A large value implies that the new data does not fit well with the correlation structure captured by the PCA model.

The second statistic was actually introduced earlier in this chapter: the Hotelling  $T^2$  statistic,

$$T_i^2 = \tilde{\mathbf{x}}_i^T \mathbf{S}^{-1} \tilde{\mathbf{x}}_i = \tilde{\mathbf{t}}_i^T \Lambda^{-1} \tilde{\mathbf{t}}_i \quad (23.59)$$

in terms of the original data, and equivalently in terms of the PCA scores and eigenvalues; it provides a measure of the variation within the new sample relative to the variation within the model. A large  $T_i^2$  value indicates that the data scores are much larger than those from which the model was developed. It provides evidence that the new data is located in a region different from one captured in the original data set used to build the PCA model. These concepts are illustrated in Fig 23.7, adapted from Wise and Gallagher, (1996)<sup>13</sup>.

To determine when large values of these statistics are significant, control limits must be developed for each one, but this requires making some distributional assumptions. Under normality assumptions, confidence limits for the  $T^2$  statistic are obtained from the  $F$ -distribution, as indicated in Eq (23.16); for  $Q$ , the situation is a bit more complicated but the limits are still easily computed numerically (see e.g., Wise and Gallagher (1996)). Points falling outside of the control limits then indicate an out-of-control multivariate process in precisely the same manner as with the univariate charts of the previous chapter. These concepts are illustrated in Fig 23.8 for process data represented with 2 principal components.

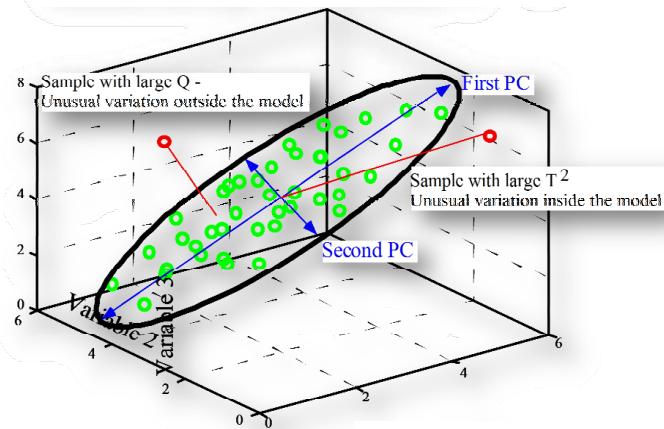
The Wise and Gallagher reference contains, among other things, additional discussions about the application of PCA in process monitoring.

### Model Building in Systems Biology

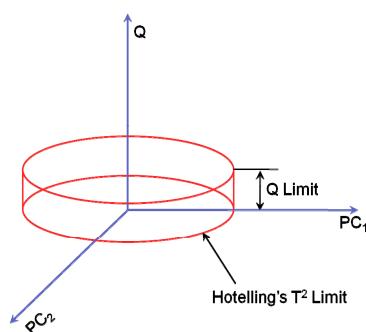
PCA continues to find application in many non-traditional areas, with its infiltration into biological research receiving increasing attention. For example, the applications of PCA in regression and in building models have been extended to signal transduction models in Systems Biology. From multivariate

---

<sup>13</sup>Wise, B.M. and N. B. Gallagher, (1996). "The process chemometrics approach to process monitoring and fault detection," *J Process Control*, 6 (6) 329–348.



**FIGURE 23.7:** Principal component model for a 3-dimensional data set described by two principal components on a plane, showing a point with a large  $Q$  and another with a large  $T^2$  value.



**FIGURE 23.8:** Control limits for  $Q$  and  $T^2$  for process data represented with two principal components.

measurements of various aspects of intracellular signalling molecules, PCA has been used to find, among other things, fundamental dimensions that appear to constitute molecular “basis axes” within the signaling network that the cell uses for the apoptotic program<sup>14</sup>. An application of PCA to the analysis of protein conformational space can also be found in Tendulkar, *et al.* (2005)<sup>15</sup>. Upon visualizing the distribution of protein conformations in the space spanned by the first four PCs as a set of conditional bivariate probability distribution plots, the peaks are immediately identified as corresponding to the preferred protein conformations.

A brief primer on PCA, especially from the perspective of analyzing molecular biology data, is available in Ringnér (2008)<sup>16</sup>. A review of other applications of PCA and PLSR in deriving biological insights from multivariate experimental data, along with a discussion of how these techniques are becoming standard tools for systems biology research, is contained in Janes and Yaffe (2006)<sup>17</sup>.

---

### 23.4 Summary and Conclusions

Strictly speaking, this final chapter in the trilogy of chapters devoted to showcasing substantial, subject-matter-level applications of probability and statistics, is different from the other two. While Chapters 21 and 22 cover true applications of probability and statistics, what this chapter covers is not so much an “application” as it is an extension to higher dimensions. In this sense, what Chapters 8–20 are to Chapter 4 (on the univariate random variable) this chapter is—albeit in condensed, miniaturized form—to Chapter 5. The coverage of multivariate probability models was brief, and even then, only those whose scalar analogs play significant roles in univariate analysis were featured. This device made it easy to place the roles of the multivariate distributions in multivariate analysis in context quickly and efficiently.

In this era of facilitated acquisition and warehousing of massive data sets, the well-trained scientist and engineer (especially those working in the chemical and other manufacturing industries) must be aware of the intrinsic nature of multivariate data, and of the random variables from which such data

<sup>14</sup>Janes, K. A., J. G. Albeck, S. Gaudet, P. K. Sorger, D. A. Lauffenburger, and M. B. Yaffe (2005): “A Systems Model of Signaling Identifies a Molecular Basis Set for Cytokine-Induced Apoptosis” *Science*, 310, 1646-1653.

<sup>15</sup>Tendulkar, A. V., M. A. Sohoni, B.A. Ogunnaike, and P. P. Wangikar, (2005). “A geometric invariant-based framework for the analysis of protein conformational space,” *Bioinformatics*, 21 (18) 3622-3628.

<sup>16</sup>Ringnér, M. (2008). “What is principal component analysis?” *Nature Biotech.* 26(3), 303-304.

<sup>17</sup>K. A. Janes and M. B. Yaffe, (2006). “Data-driven modelling of signal-transduction networks,” *Nature*, 442, 820-828.

arose—and how to analyze such data appropriately. Unfortunately, only the brief overview presented here is possible in the amount of space available in such a textbook. But our objective was never comprehensive coverage; rather it was to make the reader aware that multivariate analysis is more than merely a “multiplication” of univariate ideas by  $n$ ; probability model structures become significantly more complicated, and only the analysis techniques tailored to such complex structures can be effective for multivariate problem-solving.

In light of the limited space budget, the amount of time and space allocated to principal components analysis (PCA) might be surprising at first, but this is in recognition of how pervasive this technique is becoming (or, in some cases, has already become) in many modern manufacturing enterprises, including the traditionally conservative pharmaceutical industry. Still, a more comprehensive discussion of PCA was not possible. As such, how the approach itself was presented, and the special illustrative example employed, were jointly calibrated mostly to promote fundamental understanding of a technique that is prone to misuse and misinterpretation of its results because it is not always easy to understand.

Finally, we believe that there is no better way to demonstrate PCA and to develop an understanding of how to interpret its results than by actually *doing* real multivariate data analysis on real data sets. The project assignment at the end of the chapter offers such an opportunity to apply PCA to a real molecular biology data set.

## REVIEW QUESTIONS

1. What is a multivariate probability model? How is it related to the single variable pdf?
2. The role of the Gaussian distribution in univariate analysis is played in multivariate analysis by what multivariate distribution?
3. The Wishart distribution is the multivariate generalization of what univariate distribution?
4. Hotelling's  $T$ -squared distribution is the multivariate generalization of what univariate distribution?
5. What is the multivariate generalization of the  $F$ -distribution?
6. The Dirichlet distribution is the multivariate generalization of what univariate distribution?
7. What is Principal Components Analysis (PCA) and what is it useful for?
8. In Principal Components Analysis (PCA), what is a “loading” vector and a “score” vector?

9. How are the loading vectors in PCA related to the data matrix?
10. How are the score vectors in PCA related to the data matrix and the loading vectors?
11. In PCA, what is a Scree plot?
12. How is PCA used in process monitoring?

## PROJECT ASSIGNMENT

### Principal Components Analysis of a Gene Expression Data Set

In the Ringnér (2008) reference provided earlier, the author used a set of microarray data on the expression of 27,648 genes in 105 breast tumor samples “to illustrate how PCA can be used to represent samples with a smaller number of variables, visualize samples and genes, and detect dominant patterns of gene expression.” Only a brief summary of the resulting analysis was presented. The data set, collected by the author and his colleagues, and published in Saal, *et al.*, (2007)<sup>18</sup>, is available through the National Center for Biotechnology Information Gene Expression Omnibus database (accession GSE5325) and from [http://icg.cpmc.columbia.edu/faculty\\_parsons.htm](http://icg.cpmc.columbia.edu/faculty_parsons.htm).

Consult the original research paper, Saal, *et al.*, (2007), (which includes the application of other statistical analysis techniques, such as the Mann-Whitney-Wilcoxon test of Chapter 18, but not PCA) in order to understand the research objectives and the nature of the data set. Then download the data set and carry out your own PCA on it. Obtain the scores and loading plots for the first three principal components and generate 2-D plots similar to those in Fig 23.6 in the text. Interpret the results to the best of your ability. Write a report on your analysis and results, comparing them where possible with those in Ringnér (2008).

---

<sup>18</sup>Saal, L.H., P. Johansson, K. Holmberg, *et al.* (2007). “Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity,” *Proc. Natl. Acad. Sci. USA* 104, 7564–7569.

---

## Appendix

*Knowledge is of two kinds:  
we know a subject ourselves  
or we know where we can find information on it.*

Samuel Johnson Cowper (1709–1784)

---

### Computers and Statistical Computations

In traditional textbooks on probability and statistics, this part of the book is where one would typically find tables for generating random numbers, and for determining (tail area) probabilities for different sampling distributions. The collection usually includes tables of standard normal probabilities, or  $z$  tables;  $t$ -tables for various degrees of freedom,  $\nu$ ;  $\chi^2(\nu)$  tables;  $F$ -tables for a select combination of degrees of freedom; even binomial and Poisson probability tables. Generations of students learned to use these tables to determine probabilities, a procedure that often requires interpolation, since it is impossible to provide tables dense enough to include all possible variates and corresponding probabilities.

Things are different now. The universal availability of very fast and very powerful computers of all types—from personal desktop machines to servers and supercomputers—has transformed all aspects of scientific computing. So many software packages have now been developed specifically for carrying out every conceivable computation involved in probability and statistics that many of what used to be staples of traditional probability and statistics textbooks are no longer necessary, especially statistical tables. Random numbers can be generated for a wide variety of distributions; descriptive and graphical analysis, or estimation, regression analysis and hypothesis tests, and much more, can all now be carried out with ease. The computer has thus rendered printed statistical tables essentially obsolete by eliminating the need for them, but also by making it possible to have them available electronically. In this book, therefore, we are departing from tradition and will *not* include any statistical tables. What we are supplying here instead is a compilation of useful information about some popular software packages, and, for those who

might still want them, on-line electronic versions of statistical tables; we also include a few other on-line resources that the reader might find useful.

### Statistical Software Packages

We adopted the use of MINITAB in this text primarily because, for the uninitiated, its learning curve is *not* steep at all; its drop-down menus are intuitive, and anyone familiar with Excel spreadsheets can learn to use MINITAB with little or no instruction. Of course, several other packages are equally popular and can be used just as well. Below is a list of just a few of these, and where to find further information about them.

1. **MINITAB:** A commercial software package with a free 30-day trial download available at:  
<http://www.minitab.com/downloads/>  
A student version and academic licensing are also available.
2. **R:** A free software package for statistical computing and graphics that is based on a high level language:  
<http://www.r-project.org/> or  
<http://cran.r-project.org/>.
3. **S-Plus:** a commercial package based on the S programming language:  
<http://www.insightful.com/products/splus/>  
A free 30-day trial CD is available upon request.
4. **MATLAB Statistics Toolbox:** A commercial toolbox for use with MATLAB, the general purpose scientific computing software popular among control engineers:  
<http://www.mathworks.com/products/statistics/>
5. **SAS:** A commercial package:  
<http://www.sas.com/technologies/analytics/statistics/>
6. **JMP:** An alternative commercial product from SAS with a free 30-day trial download available at:  
<http://www.jmp.com/>  
Academic licensing is also available.
7. **STATISTICA:** A commercial package:  
<http://www.statsoft.com/products/products.htm>.

### On-line Calculators and Electronic Tables

As noted above, the advent of computer software packages has essentially eliminated the need for traditional statistical tables. Even then, these tables

are now freely available electronically on-line. Some are deployed fully in electronic form, with precise probability or variate values computed on request from pre-programmed probability distribution functions. Others are “electronic” only in the sense that the same numbers that used to be printed on paper are now made available in an on-line table; they still require interpolation. Either way, if all one wants to do is simply compute tail area probabilities for a wide variety of the usual probability distributions employed in statistical inference, a dedicated software package is not required. Below is a listing of three electronic statistical tables, their locations, and a brief summary of their capabilities.

1. <http://stattrek.com/Tables/StatTables.aspx>  
Conceived as a true on-line calculator, this site provides the capability for computing, among other things, all manner of probabilities for a wide variety of discrete and continuous distributions. There are clear instructions and examples to assist the novice.
2. <http://stat.utilities.googlepages.com/tables.htm>  
(SurfStat statistical tables by Keith Dear and Robert Brennan.)  
Truly electronic versions of  $z$ -,  $t$ -, and  $\chi^2$  “tables” are available with a convenient graphical user-interface that allows the user to specify either the variate and compute the desired tail area probabilities, or to specify the tail area probability and compute the corresponding variate. The  $F$ -tables are available only as text.
3. <http://www.statsoft.com/textbook/sttable.html>  
Provides actual tables of values computed using the commercial STATISTICA BASIC software; available tables include  $z$ -,  $t$ -,  $\chi^2$ ; the only  $F$ -tables available are for  $F(0.1, 0.05)$  and  $F(0.025, 0.01)$ . (The site includes animated Java images of the various probability distributions showing various computed, and constantly changing, cumulative probabilities.)

## Other On-line Resources

Below is a listing of other resources available on-line:

1. **Glossary:** A glossary of statistical terms arranged in alphabetical order is available at:  
[http://www.animatedsoftware.com/elearning/  
Statistics%20Explained/glossary/se\\_glossary.html](http://www.animatedsoftware.com/elearning/Statistics%20Explained/glossary/se_glossary.html)
2. **Electronic Handbooks:** Two of the most comprehensive and most useful are listed below:
  - (a) The NIST/SEMATECH e-Handbook of Statistical Methods:  
<http://www.itl.nist.gov/div898/handbook/>, and

- (b) The StatSoft Electronic Statistics Textbook:  
<http://www.statsoft.com/textbook/stathome.html>
3. **Data sets:** The following site contains links to a wide variety of statistical data, categorized by subject area and government agency. In addition, it provides links to other sites containing statistical data.  
<http://www.libraryspot.com/statistics.htm>
- Also, NIST, the National Institute of Standards and Technology, has a site dedicated to reference data sets with certified computational results. The original purpose was to enable the objective evaluation of statistical software. But instructors and students will find the data sets to be a good source of extra exercises (based on certified data).  
<http://www.itl.nist.gov/div898/strd/index.html>

---

# Index

- acceptance sampling, 222  
acceptance sampling plans, 937  
acceptance/rejection criteria, 937  
action potentials, 269, 772  
aerial bombardment of London, 869  
aggregate descriptions, 4, 409  
aggregate ensemble, 68, 100  
Albeck, J. G., 1002  
alternative hypothesis, 554  
Anderson, T. W., 771  
Anderson-Darling  
    statistic, 736  
    test, 736  
Anderson-Darling test, 771, 775  
    more sensitive than K-S test, 772  
    test statistic, 771  
ANOVA  
    fixed, random, mixed effects, 803  
ANOVA (Analysis of Variance), 311, 605,  
    796  
    central assumptions, 796  
    identity, 672, 800  
ANOVA tables, 675  
    one-way classification, 801  
Antoine's equation, 728  
arithmetic average, 14  
Asprey, S., 839  
Assisted Reproductive Technology (ART),  
    364  
Atkinson, A., 839  
Atwood, C.L., 251  
automatic process control, *see* engineering process control  
Avandia®, 138, 141  
average deviation, 440  
Balakrishnan, N., 979  
bar chart, 419  
bar graph, *see* bar chart  
baroreceptors, 929  
Bayes' estimator, 520  
Bayes' rule, 76  
Bayes' Theorem, 519  
Bayes, Revd. Thomas, 77, 519  
Bayesian estimation, 518–527, 862  
    recursive, 525  
        application to Prussian army data, 860  
Bequette, B. W., 976  
Bernoulli random variable, 221, 761  
    mathematical characteristics, 222  
    probability model, 221  
Bernoulli trial, 221, 225, 372  
Bernoulli, Jakob, 198  
beta distribution, 522, 983  
    many shapes of, 303  
beta random variable, 301–308  
    application, 305  
        in quality assurance, 305  
    generalized, 307  
    inverted, 308  
    mathematical characteristics, 302  
    probability model, 302  
    relation to gamma random variable, 301  
Bibby, J. M., 982  
binary classification tests, 558  
    sensitivity, 559  
    specificity, 559  
binomial distribution, 761  
binomial random variable, 225–230, 280,  
    372  
    applications, 227  
    inference, 228  
    mathematical characteristics, 226  
    probability model, 226  
    relation to other random variables, 227  
Birtwistle, M. R., 269, 330, 453, 839  
Bisgaard, S., 851

- bivariate Gaussian distribution, 980  
 bivariate random variable, 139  
   continuous, 149  
   definition, 139  
   discrete, 150  
   informal, 140  
 blocking, 805  
 blood pressure control system, mammalian, 929  
 Boltzmann, L. E., 353  
 Borel fields, 68  
 box plot, 427, 429  
 Box, G. E. P., 731, 811, 834, 851  
 Brauer, F., 890  
 Burman, J. P., 832
- C-chart, 957  
 calculus of variations, 345  
 calibration curves, 659  
 Cardano, 198  
 Castillo-Chavez, C., 890  
 catalyst, 12  
 Cauchy distribution, 786  
   application in crystallography, 789  
   model for price fluctuations, 786  
 Cauchy random variable, 314  
   application  
     high-resolution price fluctuations, 316  
   mathematical characteristics, 315  
   probability model, 314  
   relation to other random variables, 316  
 cell culture, 173  
 Central Limit Theorem, 288, 468, 471, 608, 947  
 chain length distributions, 235  
   most probable, 235  
 Chakraborti, S, 778  
 characteristic function, 115–116, 177  
   inversion formula, 116  
 characteristic parameters, 409  
 Chebyshev's inequality, 121, 229, 492  
 chemical engineering  
   illustration, 35  
   principles, 38  
 chemical process control, 964–969  
 chemical reactors, 35  
 chi-square random variable  
   application, 272  
   special case of gamma random variable, 271  
 Chi-squared goodness-of-fit test, 739–745  
   relation to *z*-test, 745  
 chi-squared test, 601  
 Chinese Hamster Ovary (CHO) cells, 330, 361, 453  
 Clarke, R. D., 868  
 co-polymer composition, 140  
 coefficient of determination  
   adjusted,  $R_a^2 dj$ , 673  
 coefficient of determination,  $R^2$ , 672, 673  
 coefficient of variation, 109  
 commercial coating process, 879  
 completely randomized design, 798  
   balanced or unbalanced, 799  
 complex system  
   crosslinks, 909  
   series-parallel configuration, 906  
 complex variable, 115  
 computer programs, 426  
 conditional distribution  
   general multivariate, 153  
 conditional expectation, 156  
   bivariate, 156  
 conditional mean, 157  
 conditional probability  
   empirical, 209  
 conditional probability distribution  
   definition, 147  
 conditional variance, 157  
 confidence interval, 509  
   around regression line, 668  
   relation to hypothesis tests, 575  
 confidence interval, 95%, 507  
   mean of normal population, 507  
   on the standard deviation, 510  
 confidence intervals  
   in regression, 661  
 conjugate prior distribution, 862  
 consistent estimator, 492  
 constrained least squares estimate, 696  
 continuous stirred tank reactor (CSTR), 35, 193, 261, 325, 361  
 control charts, 946  
   C-chart, 957  
   CUSUM charts, 961  
   EWMA charts, 963

- graphical means of testing hypotheses, 946  
I and MR chart, 952  
P-chart, 954  
S-chart, 950  
Shewhart chart, 947  
Western Electric Rules, 960  
Xbar-R chart, 951  
Xbar-S chart, 950
- control hardware electronics, 143  
convolution integrals, 179  
Corbet, S., 255  
correlation coefficient, 158  
Coughlin, R., 210  
covariance, 157  
covariance matrix, 689  
Cramér-Rao inequality, 492  
Cramér-Rao lower bound, 492  
critical region, 556  
cumulative distribution function, 96, 99, 100  
bivariate, 142  
cumulative hazard function, 124, 272  
CUSUM charts, 961
- Dantzig, T., 363  
Darling, D. A., 771  
data  
nominal, 418  
ordinal, 418  
data characteristics, 436–442  
central location, 436  
variability, 440  
de Laplace, Pierre-Simon, 279  
de Moivre, Abraham, 198, 279  
death-by-horse kicks, 858  
deaths in London, 41  
DeMorgan's Laws, 62  
Desemone, J., 976  
design of experiments, 415  
deterministic, 14  
expression, 34  
idealization, 2  
phenomena, 3  
differential entropy, 342  
Dirac delta function, 38  
Dirichlet distribution, 983  
Dirichlet, J. P. G. L., 983  
discrete uniform random variable, 219  
application, 220  
model, 220
- distribution  
conditional, 147  
joint, 141  
joint-marginal, 156  
marginal, 144, 145, 156  
multimodal, 117  
posterior, 520  
prior, 520  
symmetric, 109  
unimodal, 117
- distributions, 95, 107  
leptokurtic, 111  
moments of, 107  
of several random variables, 141  
platykurtic, 111  
relationship between joint, conditional, marginal, 519
- DNA replication origins  
distances between, 269
- Donev, A., 839  
Doyle, III F. J., 976  
Draper, N. R., 834
- economic considerations, 12  
efficiency, 491  
efficient estimator, 491  
empirical frequencies, 205  
engineering process control, 966  
feedback error, 966  
Proportional-Integral-Derivative (PID) controllers, 966
- engineering statistics, 2  
ensemble, 41  
entropy, 119, 338–344  
of Bernoulli random variable, 340  
of continuous random variable, 340  
differential entropy, 342  
of deterministic variable, 339  
of discrete uniform random variable, 339  
Erlang distribution, *see* gamma distribution
- Erlang, A. K., 266  
estimation, 489  
estimator, 489  
estimator characteristics  
consistent, 492

- efficient, 491
- unbiased, 490
- estimators
  - criteria for choosing, 490–493
  - method of moments, 493
  - sequence of, 492
- Euler equations, 345, 346, 348–350
- event, 59
  - certain, 61, 63
  - complex, compound, 60
  - impossible, 61, 63, 67
  - simple, elementary, 60
- events
  - compound, 64
  - elementary, 64
  - mutually exclusive, 63, 69, 78
- EWMA charts, 961
  - related to Shewhart and CUSUM charts, 963
- expected value, 102
  - definition, 105
  - properties
    - absolute convergence, 105, 106
    - absolute integrability, 105, 106
- experiment, 58
  - conceptual, 58, 64
- experimental studies
  - phases of, 794
- exponential distribution
  - “memoryless” distribution, 263
  - discrete analog of geometric distribution, 261
- exponential pdf
  - application in failure time modeling, 913
- exponential random variable, 260–264
  - applications, 263
  - mathematical characteristics, 262
  - probability model, 261
  - special case of gamma random variable, 266
- extra cellular matrix (ECM), 361
- F distribution, 311, 474
- F random variable, 309–311
  - application
    - ratio of variances, 311
  - definition, 309
  - mathematical characteristics, 310
- probability model, 310
- F-test, 604
  - in regression, 674
- sensitivity to normality assumption, 605
- factor levels, 795
- factorial designs,  $2^k$ , 814
  - characteristics
    - balanced, orthogonal, 816
    - model, 816
    - sample size considerations, 818
  - factorial experiments, 814
  - factors, 795
  - failure rate, 911
  - failure times
    - distribution of, 913
  - failure-rate
    - decreasing, model of, 914
    - increasing, model of, 914
  - Fermat, 198
  - first order ODE, 38
  - first-principles approach, 2
  - Fisher information matrix, 837, 848
  - Fisher, R. A., 209, 255, 309, 541, 858
  - fluidized bed reactor, 325
  - Fourier transforms, 116
  - fractional factorial designs
    - alias structure, 824
    - defining relation, 824
    - design resolution, 824
    - folding, 826
    - projection, 825
  - frequency distribution, 18, 20, 424, 426
  - frequency polygon, 426
  - frequentist approach, 426
  - functional genomics, 305
  - Gallagher, N. B., 1000
  - gamma distribution, 266
    - generalization of Erlang distribution, 266
    - model for distribution of DNA replication origins, 269
  - gamma pdf
    - application in failure time modeling, 917
  - gamma random variable, 180, 181, 264–271, 462
    - applications, 269

- generalization of exponential random variable, 264  
mathematical characteristics, 266  
probability model, 265  
reproductive property of, 181
- Garge, S., 853
- Gaudet, S., 1002
- Gauss, J.C.F., 279
- Gaussian distribution, 654  
bivariate, 980
- Gaussian probability distribution, 288
- Gaussian random variable, 279–292  
applications, 290  
basic characteristics, 288  
Herschel/Maxwell model, 285–287  
limiting case of binomial random variable, 280  
mathematical characteristics, 288  
misconception of, 288  
probability model, 288  
random motion in a line, 282–285
- Gelmi, C. A., 306
- gene expression data set, 1004
- genetics, 199
- geometric random variable, 234  
application, 235  
mathematical characteristics, 234  
probability model, 234
- geometric space, 44
- Gibbons, J. D., 778
- Gossett, W. S., 312
- Gram polynomials, 706
- granulation process, 296, 297
- graphical techniques, 442
- Graunt, John, 41
- gravitational field, 44
- Greenwood, M., 252
- group classification, 18
- hazard function, 123, 272, 911  
bathtub curve, 912  
equivalently, failure rate, 912
- heat transfer coefficient, 34
- Hendershot, R. J., 837
- hereditary factors, 203  
genes, 203
- heredity  
dominance/recessiveness, 203
- genotype, 204
- phenotype, 204
- Heusner, A. A., 729
- Hirschfelder-Curtiss-Bird, 638
- histogram, 18, 424  
for residual errors, 678
- Hoerl, A.E., 697
- Hotelling  $T^2$  statistic, 1000
- Hotelling's  $T$ -squared distribution, 982
- Hotelling, H., 982
- housekeeping genes, 764
- Hunter, J. S., 811
- Hunter, W. G., 811
- Huygens, 198
- hydrocarbons, 455
- hypergeometric random variable, 222  
application, 224  
mathematical characteristics, 224  
probability model, 223
- hypothesis  
alternative,  $H_a$ , 560  
null,  $H_0$ , 560  
one-sided, 555  
two-sided, 554
- hypothesis test, 555  
 $p$ -value, 560  
error  
Type I, 557  
Type II, 558  
general procedure, 560  
non-Gaussian populations, 613  
power and sample size determination, 591–600  
power of, 558  
risks  
 $\alpha$ -risk, 558  
 $\beta$ -risk, 558  
two proportions, 611  
using MINITAB, 573
- hypothesis test, significance level of, 557
- hypothesis testing  
application to US census data, 876  
in regression, 664
- ideal probability models, 4
- in-vitro fertilization (IVF), 42, 101, 225, 413
- binomial model  
sensitivity analysis of, 393
- binomial model for, 372, 377

- binomial model validation, 375
- Canadian guidelines, 370
- central characteristics, 371
- clinical data, 367
- clinical studies, 367
- Elsner clinical data, 375
- Elsner clinical study, 369
- Elsner study
  - study characteristics, 375
  - guidelines and policy, 370
  - implantation potential, 369, 372
  - mixture distribution model, 382
  - model-based optimization, 384
  - multiple births, 365
    - risk of, 366
  - oocyte donation, 364
  - optimization problem, 385
  - optimum number of embryos,  $n^*$ , 385, 387
  - patient categorization, 390
  - SEPS parameter
    - non-uniformity, 380
  - single embryo probability of success parameter (SEPS), 372
  - treatment cycle, 372
  - treatment outcomes
    - theoretical analysis of, 390
- in-vitro fertilization (IVF) treatment
  - binomial model of, 397
  - model-based analysis of, 393
- inclusions, 16
- independence
  - pairwise, 78
  - stochastic, 158
- information
  - quantifying content of, 337
  - relation to uncertainty, 336
- information content, 119
- interspike intervals
  - distribution of, 772
- interval estimate, 490
- interval estimates, 506–518
  - difference between the two population means, 512
  - for regression parameters, 661
  - mean,  $\sigma$  unknown, 508
  - non-Gaussian populations, 514
  - variance of normal population, 510
- interval estimation, 489
- interval estimator, 490
- inverse bivariate transformation, 182
- inverse gamma random variable, 325, 550
- inverse transformation, 172
- Jacobian
  - of bivariate inverse transformation, 182
  - of inverse transformation, 175, 176
- Janes, K. A., 1002
- Johnson, N. L., 979
- joint probability distribution
  - definition, 142
- joint probability distribution function, 144
- Jovanovic, L., 976
- Kalman filter, 700
- kamikaze pilots, 209
- Kent, J. T., 982
- keystone component, 910
- Kholodenko, B. N., 839
- Kimball, G. E., 210
- Kingman, J.F.C., 68
- Kleiber's law, 194, 729
- Kolmogorov, 67, 98, 198
- Kolmogorov-Smirnov (K-S) test, 770
  - test statistic, 771
- Konishi, S., 970
- Kotz, S., 979
- Kruskall-Wallis test, 805
  - nonparametric one-way ANOVA, 805
- kurtosis, 111
  - coefficient of, 111
- Lagrange multiplier, 345, 346, 353
- Lagrangian functional, 345
- Laplace, 198, 519
- Laplace transforms, 116
- Lauffenburger, D. A., 1002
- Lauterbach, J., 837
- law of large numbers, 229
- least squares
  - estimator
    - properties, 660
    - unbiased, 660
  - least squares constrained, 696

- estimator, 652  
method of, 653  
ordinary, 654  
principles of, 651, 652  
recursive, 698, 699  
weighted, 694
- Lenth, R. V., 829
- life tests, 919  
    accelerated tests, 919  
    nonreplacement tests, 919  
    replacement tests, 919  
    test statistic, 920  
    truncated tests, 919
- life-testing, 275
- likelihood function, 497
- likelihood ratio tests, 616–623
- linear operator, 107
- log-likelihood function, 498
- logarithmic series distribution, 255
- logarithmic series random variable, 248
- logistic distribution, 329  
    (standard), 326
- lognormal distribution, 293  
    multiplicative characteristics, 294
- lognormal random variable, 292–297  
    applications, 296  
    central location of  
        median more natural, 296  
    mathematical characteristics, 293  
    probability model, 293  
    relationship to Gaussian random  
        variable, 293
- loss function  
    quadratic, 971
- Lucas, J. M., 482, 750, 963
- Macchietto, S., 839
- macromolecules, 113
- Malaya, butterflies of, 255, 541
- Mann-Whitney  $U$  test statistic, 767
- Mann-Whitney-Wilcoxon (MWW) test,  
    766–769  
    nonparametric alternative to 2-  
        sample  $t$ -test, 769
- manufactured batch, 66
- Mardia, K. V., 982
- marginal expectation, 156
- marginal probability distribution  
    definition, 145  
    marginal variance, 156, 157  
Markov Chain Monte Carlo (MCMC),  
    527
- Markov's inequality, 121
- Marquardt, D.W., 698
- material balance, 38
- mathematical biology, 209
- mathematical expectation, 102, 154  
    marginal, 156  
    of jointly distributed random vari-  
        ables, 154
- maximum *a-posteriori* (MAP) estimate,  
    863
- maximum *a-posteriori* (MAP) estimator,  
    520
- maximum entropy distribution, 346  
    beta pdf, 351, 354  
    continuous uniform pdf, 348  
    discrete uniform pdf, 352  
    exponential pdf, 349, 352  
    gamma pdf, 354  
    Gaussian pdf, 350, 352  
    geometric pdf, 347, 352
- maximum entropy models, 344–354  
    from general expectations, 351
- maximum entropy principle, 344
- maximum likelihood, 496–503
- maximum likelihood estimate, 498  
    characteristics, 501  
    Gaussian population parameters,  
        500  
    in regression, 657
- maximum likelihood principle, 616
- mean, 437  
    limiting distribution of, 467  
    sample, 438  
    sampling distribution of, 465  
        normal approximation, 468  
        standard error of, 467
- mean absolute deviation from the me-  
    dian (MADM), 440
- mean-time-between-failures (MTBF), 913
- median, 117, 118, 437  
    sample, 439
- melt index, 140
- Mendel's experiments  
    multiple traits  
        pairwise, 205
- pea plants

- characteristics, 199
- genetic traits, 200
- results, 207
- single traits, 201
- Mendel, Gregor, 199
- method of moments, 493–496
- method of moments estimators
  - properties
    - consistent, 496
    - not unique, 496
- microarray
  - reference spot, 194
  - test spot, 193
- microarray data, 306
  - fold change ratio, 194
- mixture distributions
  - Beta-Binomial, 328, 382
    - application to Elsner data, 400
  - Poisson-Gamma, 278
- mode, 117, 438
  - sample, 439
- molecular weight
  - $z$  average,  $M_z$ , 113
  - distributions (MWD), 113
  - non-uniform, 113
  - number average,  $M_n$ , 113
  - weight average,  $M_w$ , 113
- molecular weight distribution, 140
- moment generating function, 113
  - independent sums, 115
  - linear transformations, 115
  - marginal, 156
  - uniqueness, 114
- moments, 107
  - $k^{th}$  ordinary, 107
  - central, 108
  - first, ordinary, 107
  - second central, 108
- monomer molecules, 41
- Mooney viscosity, 952
- Morse, P. M., 210
- multinomial random variable, 231
- multivariate normal distribution, 980
- multivariate probability model, 978
- multivariate process monitoring, 999
- multivariate random variable
  - definition, 141
- multivariate transformations, 184
  - non-square, 185
- overdefined, 185
- underdefined, 185
- square, 184
- Mylar®, 192
- negative binomial distribution, 233
  - as the Poisson-Gamma mixture distribution, 278
- negative binomial random variable, 232–234
  - mathematical characteristics, 233
  - probability model, 232
    - alternative form, 233
- Nelson, W. B., 919
- nonparametric methods, 759
  - robust, 759
- normal approximation, 468
- normal distribution, *see* Gaussian probability distribution
- normal equations, 655
  - matrix form of, 689
- normal population, 471
- normal probability plot, 735
  - application in factorial designs, 829
- normality test
  - for residual errors, 678
- null hypothesis, 554
- Ogunnaike, B. A., 618, 721, 728, 837, 839, 929, 952, 966, 1002
- one-sample sign test, 760
  - nonparametric alternative to one-sample *t*-test, 763
  - nonparametric alternative to one-sample *z*-test, 763
    - test statistic, 761
    - sampling distribution, 761
- operating characteristic curve, 939
  - relation to power and sample size, 942
- opinion pollster, 488
- optimal experimental designs, 837
  - A-, D-, E-, G-, and V-, 838
- Ospina, P., 976
- outcome, 58
- outcomes
  - equiprobable, 76
- outlier, 428
- overdispersion, 242

- Owens, C., 976
- P-chart, 954
- p-value, 560
- boderline, 623
  - observed significance level, 560
- Pólya, George, 233
- paper helicopter, 891
- Pareto chart, 421
- Pareto random variable, 328
- Pareto, V. F., 421
- particle size distribution, 113, 296
- particulate products, 113
- parts, defective, 66
- Pascal distribution, 233
- Pascal, Blaise, 198, 233
- pdf, 100
- Pearson, 18
- Pearson, K., 985
- percentiles, 119
- phenomenological mechanisms, 3
- Philadelphia Eagles, 785
  - point differential, 2008/2009 season, 785
  - points scored by, 2008/2009 season, 785
- pie chart, 421
- Plackett, R.L., 832
- plug flow reactor (PFR), 35
- point estimate, 489
- point estimates
- precision of, 503–506
    - binomial proportion, 505
    - mean,  $\sigma$  known, 504
    - mean,  $\sigma$  unknown, 505
    - variance, 505
  - point estimation, 489
  - Poisson distribution, 496
  - Poisson events, 260
  - Poisson model, 859
  - Poisson random variable, 173, 174, 236–243, 463
    - applications, 240
    - limiting form of binomial random variable, 236
    - mathematical characteristics, 239
    - overdispersed, 242
      - negative binomial model, 242
    - probability model, 237, 239
- probabilty model
- from first principles, 237
- Poisson-Gamma mixture distribution, 276–278
- Polya distribution, *see* negative binomial distribution
- polydispersity index (PDI), 113
- polymer reactor, 143
- polymer resin, 63
- polymeric material, 113
- polymerization
- free radical, 235
- polynomial regression, 701
  - orthogonal, 704
- population, 411, 488
  - dichotomous, 222
- posterior distribution, 520, 522, 863
- Pottmann, M., 618, 929
- power and sample size
- computing with MINITAB, 599
- pre-image, 91
- prediction error, 668
- prediction intervals, 668
- principal component
- loading vectors, 986
  - score vectors, 987
- principal components, 985
- principal components analysis (PCA), 985
  - application in systems biology, 1000
  - scale dependent, 986
  - Scree plot, 990
- prior distribution, 520
  - uniform, 523
- probabilistic framework, 3, 47
- probability, 43, 69
  - à-posteriori*, 77
  - à-priori*, 77
  - application by Mendel, 204
  - bounds, 119
    - general lemma, 120
  - calculus of, 68, 69
  - classical *à-priori*, 45
  - conditional, 72–74, 147
  - equiprobale assignment of, 70
  - mathematical theory of, 67
  - relative frequency *à-posteriori*, 46
  - set function, 67, 90
  - bivariate, 139

- induced, 90, 95
- subjective, 46
- theory, 414
- total, 74
  - Theorem of, 76
- probability density function
  - definition, 98
  - joint, 142
- probability distribution function, 23, 43, 47, 78, 96, 100
  - conditional, 147
  - convolution of, 179
  - definition, 98
  - joint, 142
  - marginal, 145
- probability distributions
  - chart of connections, 319
- probability model validation, 732
- probability paper, 734
- probability plots, 733–739
  - modern, 736
- probability theory, 58
  - application to in-vitro fertilization, 395
- probability
  - total
    - Theorem of, 910
- process
  - chemical, 12
  - chemical manufacturing, 2
  - manufacturing, 16, 43, 71
  - yield, 12
- process control, 410, 944
- process dynamics, 275, 410
- process identification, 410
- product law of reliabilities, 903
- product law of unreliabilities, 904
- product quality, 16
- propagation-of-errors
  - application, 194
- Q statistic, 1000
- quality assurance, 16
- quantiles, 119
- quantization error, 342
- quartiles, 118
- random components, 11
- random fluctuations, 205
- random mass phenomena, 41
- random phenomena, 3
- random sample, 460
- random variability, 14
- random variable, 90, 103, 412
  - n*-dimensional, 141, 172
  - bivariate, 139
    - continuous, 146
  - continuous, 94
  - definition, 90
  - discrete, 94
  - entropy
    - definition, 119
  - entropy of, 338
  - informal, 94
  - kurtosis, 111
  - mechanistic underpinnings of, 218
  - moments
    - practical applications, 113
  - multivariate, 164, 978
  - ordinary moment of, 107
  - realizations, 409
  - skewness, 109
  - two-dimensional, 95, 143
- random variable families
  - Gamma family, 259
  - generalized model, 276
  - Gaussian family, 278
    - generalized model, 300
  - Ratio family, 300
- random variable space, 96, 103
  - bivariate, 139
- random variable sums
  - pdfs of, 177
    - cdf approach, 177
    - characteristic function approach, 177, 180
- random variable transformations
  - bivariate, 182
  - continuous case, 175
  - discrete case, 173
  - general continuous case, 176
  - non-monotone, 188
  - single variable, 172
- random variables
  - mutually stochastically independent, 161
  - continuous, 98
  - discrete, 95

- inter-related, 139  
negatively correlated, 158  
positively correlated, 158  
randomized complete block design, 805  
Ray, W. H., 728, 952, 966  
Rayleigh distribution, 298  
relationship to Weibull distribution, 299  
Rayleigh random variable, 297  
application, 300  
probability model, 298  
regression  
multiple linear, 686  
matrix methods, 688  
regression line, 663  
regression model  
mean-centered, 677  
one-parameter, 653  
two-parameter, 653  
regression parameters, 661  
rejection region, 556  
relative frequency, 18, 424  
relative sensitivity function, 394  
reliability, 225, 900  
component, 901  
definition, 900  
system, 901  
reliability function, 911  
residence time, 35  
residence time distribution, 193  
exponential, 38  
residual  
standardized, 679  
residual analysis, 678–682  
residual error, 658, 678  
residual sum-of-squares, 662  
response, 795  
response surface designs, 834  
application to process optimization, 879  
Box-Behnken, 835  
face centered cube, 835  
ridge regression, 697  
Riemann integral, 98  
Ringnér, M., 1002  
Rogers, W. B., 837  
Ross, P.J., 970  
Ryan, T.P., 837  
S-chart, 948  
sample, 411  
sample range, 440  
sample space, 59, 61, 68, 103, 412  
discrete, 59, 69  
sample variance, 441  
sampling distribution, 462  
of single variance, 473  
of two variances, 474  
scatter plot, 431, 648  
Schwaber, J. S., 618, 929  
screening designs  
fractional factorial, 822  
Plackett-Burman, 833  
sensitivity function, 684  
set  
empty, null, 61  
set function  
additive, 66, 67  
probability, 67  
set functions, 65  
set operations, 61  
sets, 61  
algebra, 61  
complement, 61  
disjoint, 63, 66  
intersection, 61  
partitioned, 75  
union, 61  
Sherwood, T.K., 724  
Shewhart, W. A., 947  
sickle-cell anemia, 252  
signal-to-noise ratio, 597  
signed ranks, 763  
simple system, 901  
parallel configuration, 904  
series configuration, 903  
single factor experiments, 797–811  
completely randomized design, 798  
data layout, 798  
model and hypothesis, 798  
two-way classification, 805  
randomized complete block design, 805  
single proportions, exact test, 610  
skewness, 109  
coefficient of, 109  
Snively, C. M., 837  
Sohoni, M. A., 1002

- Sorger, P. K., 1002  
 specification limits, 946  
     different from control limits, 946  
 standard Cauchy distribution, 315  
 standard deviation, 109  
     pooled, 579  
     sample, 441  
 standard normal distribution, 290, 468, 471  
 standard normal random variable, 290  
     mathematical characteristics, 292  
     relationship to the Chi-square random variable, 292  
 standard uniform random variable, 308  
 statistic, 461  
 statistical hypothesis, 554  
 statistical inference, 412, 470  
     in life testing, 919  
 statistical process control, 944  
 statistics, 415  
     descriptive, 415  
     graphical, 416  
     numerical, 416  
     inductive, 415  
     inferential, 415  
 Stirzaker, D., 130  
 stochastic independence, 162  
     definition, 160  
     mutual, 163  
     pairwise, 163  
 Student's *t* random variable, 311–314  
     application, 314  
     definition, 312  
     mathematical characteristics, 312  
     probability model, 312  
 Student's *t*-distribution, 471  
 sum-of-squares function, 653  
 survival function, 122, 911  
 system reliability function, 918  
 t-distribution, 312, 314, 471, 508  
 t-test, 570  
     one-sample, 571  
     paired, 586  
     two-sample, 579, 581  
 Taguchi, G., 969  
 Taylor series approximation, 195  
 Taylor series expansion, 114  
 Taylor, S.J., 68  
 temperature control system reliability, 143  
 Tendulkar, A. V., 1002  
 test statistic, 556  
 theoretical distribution, 22  
 thermal conductivity, 455  
 thermocouple calibration, 194  
 Thermodynamics, 44  
 time-to-failure, 269, 275  
 Tobias, R., 839  
 total quality management, 935  
 transformations  
     monotonic, 175  
     non-monotonic, 176  
     nonlinear, 174  
     single variable, 172  
 treatment, 795  
 trial, 59  
 trinomial random variable, 230  
     probability model, 230  
 Tukey, J., 427  
 two-factor experiments, 811  
     model and hypothesis, 812  
     randomized complete block design, two-way crossed, 812  
 two-sample tests, 576  
 unbiased estimator, 490  
 uncertainty, 11  
 uniform random variable, 176  
 uniform random variable (continuous)  
     application  
         random number generation, 309  
     mathematical characteristics, 308  
     probability model, 308  
     relation to other random variables, 309  
     special case of beta random variable, 308  
 universal set, 62  
 US legal system  
     like hypothesis test, 556  
 US population, 435  
     age distribution, 456  
 US Population data, 870  
 variability, 2  
     common cause, special cause, 945  
 variable

- dependent, 650  
discrete, 17  
qualitative, 417  
quantitative, 417  
variable transformation  
  in regression, 685  
variables  
  dependent, 795  
  independent, 795  
  nuisance, 811  
variance  
  definition, 108  
  sample, 441  
  sampling distribution of, 473  
Venn diagram, 66  
von Bortkiewicz, L., 857
- Wall Street Journal (WSJ), 364  
Wangikar, P. P., 1002  
Weibull pdf  
  application in failure time modeling, 914  
Weibull random variable, 272–275  
  applications, 275  
  mathematical characteristics, 274  
  probability model, 273  
  relation to exponential random variable, 272  
Weibull, Waloddi, 273  
weighted average, 75  
Welf, E. S., 361  
Westphal, S. P., 364  
Wilcoxon signed rank test, 763  
  normal approximation not recommended, 764  
  restricted to symmetric distributions, 763  
Wilks Lambda distribution, 982  
Wilks, S. S., 982  
Wise, B.M., 1000  
Wishart distribution, 981  
  multivariate generalization of  $\chi^2$  distribution, 981  
Wishart, J., 981  
World War II, 209
- Xbar-R chart, 951
- Yaffe, M. B., 1002
- yield improvement, 12  
Yule, G. U., 252
- z-score, 290, 564  
z-shift, 592  
z-test, 563  
  one-sample, 566  
  single proportion, large sample, 608  
  two-sample, 577  
Zisser, H., 976  
Zitarelli, D. E., 210