GLOBAL EDITION

Probability & Statistics

for Engineers & Scientists

NINTH EDITION

Walpole • Myers • Myers • Ye

ALWAYS LEARNING

PEARSON

Probability & Statistics for Engineers & Scientists

NINTH EDITION
GLOBAL EDITION

Ronald E. Walpole
Roanoke College

Raymond H. Myers
Virginia Tech

Sharon L. Myers
Radford University

Keying Ye
University of Texas at San Antonio

Boston Columbus Hoboken Indianapolis New York San Francisco
Amsterdam Cape Town Dubai London Madrid Milan Munich Paris Montr´eal Toronto Delhi Mexico
City S˜ao Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo

Editorial Director: *Chris Hoag*
Editor in Chief: *Deirdre Lynch*
Acquisitions Editor: *Patrick Barbera*
Assistant Acquisitions Editor, Global Edition:
Aditee Agarwal
Editorial Assistant: *Justin Billing*
Program Manager: *Chere Bemelmans*
Project Manager: *Christine Whitlock*
Program Management Team Lead: *Karen Wernholm*
Project Management Team Lead: *Peter Silvia* Project
Editor, Global Edition: *K.K. Neelakantan* Senior
Manufacturing Controller, Global Edition: *Trudy Kimber*

Pearson Education Limited
Edinburgh Gate
Harlow
Essex CM20 2JE
England

and Associated Companies throughout the world
Media Production Manager, Global Edition:
Vikram Kumar
Media Producer: *Aimee Thorne*
MathXL Content Manager: *Bob Carroll*

Product Marketing Manager: *Tiffany Bitzel*
Field Marketing Manager: *Evan St. Cyr*
Marketing Coordinator: *Brooke Smith*
Senior Author Support/Technology Specialist: *Joe Vetere*
Rights and Permissions Project Manager: *Gina Cheselka*
Procurement Specialist: *Carol Melville*
Associate Director of Design USHE EMSS/HSC/EDU: *Andrea Nix*
Program Design Lead: *Heather Scott*
Cover Image: Chill Chillz / Shutterstock.com

Visit us on the World Wide Web at: www.pearsonglobaleditions.com

*Authorized adaptation from the United States edition, entitled Probability & Statistics for Engineers & Scientists,*9th *Edition MyStatLab Update, ISBN 978-0-13-411585-6, by Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, and Keying Ye published by Pearson Education* c *2017.*

Acknowledgements of third party content appear on page 18, which constitutes an extension of this copyright page.

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

10 9 8 7 6 5 4 3 2 1

ISBN 10: 1292161361
ISBN 13: 9781292161365

Typeset by Aptara
Printed and bound in Italy by LEGO

This book is dedicated to

Billy and Julie

R.H.M. and S.L.M.

Limin, Carolyn and Emily

K.Y.

This page intentionally left blank

Contents

Preface

General Approach and Mathematical Level

Our emphasis in creating this edition is less on adding new material and more on providing clarity and deeper understanding. This objective was accomplished in part by including new end-of-chapter material that adds connective tissue between chapters. We affectionately call these comments at the end of the chapter "Pot Holes." They are very useful to remind students of the big picture and how each chapter fits into that picture, and they aid the student in learning about limitations and pitfalls that may result if procedures are misused. A deeper understanding of real-world use of statistics is made available through class projects, which were added in several chapters. These projects provide the opportunity for students alone, or in groups, to gather their own experimental data and draw inferences. In some cases, the work involves a problem whose solution will illustrate the meaning of a concept or provide an empirical understanding of an important statistical result. Some existing examples were expanded and new ones were introduced to create "case studies," in which commentary is provided to give the student a clear understanding of a statistical concept in the context of a practical situation.

In this edition, we continue to emphasize a balance between theory and appli cations. Calculus and other types of mathematical support (e.g., linear algebra) are used at about the same level as in previous editions. The coverage of an alytical tools in statistics is enhanced with the use of calculus when discussion centers on rules and concepts in probability. Probability distributions and sta tistical inference are highlighted in Chapters 2 through 10. Linear algebra and matrices are very lightly applied in Chapters 11 through 15, where linear regres sion and analysis of variance are covered. Students using this text should have had the equivalent of one semester of differential and integral calculus. Linear algebra is helpful but not necessary so long as the section in Chapter 12 on multiple linear regression using matrix algebra is not covered by the instructor. As in previous editions, a large number of exercises that deal with real-life scientific and engineering applications are available to challenge the student. The many data sets associated with the exercises are available for download from the website http://www.pearsonglobaleditions.com/Walpole or in MyStatLab.

Summary of Changes

- We've added MyStatLab, a course management systems that delivers proven results in helping individual students succeed. MyStatLab provides engaging experiences that personalize, stimulate, and measure learning for each student.

  To learn more about how MyStatLab combines proven learning applications with powerful assessment, visit www.mystatlab.com or contact your Pearson representative.

- Class projects were added in several chapters to provide a deeper understand ing of the real-world use of statistics. Students are asked to produce or gather their own experimental data and draw inferences from these data.
- More case studies were added and others expanded to help students understand the statistical methods being presented in the context of a real-life situation.
- "Pot Holes" were added at the end of some chapters and expanded in others. These comments are intended to present each chapter in the context of the big picture and discuss how the chapters relate to one another. They also provide cautions about the possible misuse of statistical techniques MSL bullet.
- Chapter 1 has been enhanced to include more on single-number statistics as well as graphical techniques. New fundamental material on sampling and experimental design is presented.
- Examples added to Chapter 8 on sampling distributions are intended to motivate *P*-values and hypothesis testing. This prepares the student for the more challenging material on these topics that will be presented in Chapter 10.
- Chapter 12 contains additional development regarding the effect of a single re gression variable in a model in which collinearity with other variables is severe.
- Chapter 15 now introduces material on the important topic of response surface methodology (RSM). The use of noise variables in RSM allows the illustration of mean and variance (dual response surface) modeling.
- The central composite design (CCD) is introduced in Chapter 15.
- More examples are given in Chapter 18, and the discussion of using

Bayesian methods for statistical decision making has been enhanced.

Content and Course Planning

This text is designed for either a one- or a two-semester course. A reasonable plan for a one-semester course might include Chapters 1 through 10. This would result in a curriculum that concluded with the fundamentals of both estimation and hy pothesis testing. Instructors who desire that students be exposed to simple linear regression may wish to include a portion of Chapter 11. For instructors who desire to have analysis of variance included rather than regression, the one-semester course may include Chapter 13 rather than Chapters 11 and 12. Chapter 13 features one factor analysis of variance. Another option is to eliminate portions of Chapters 5 and/or 6 as well as Chapter 7. With this option, one or more of the discrete or con tinuous distributions in Chapters 5 and 6 may be eliminated. These distributions include the negative binomial, geometric, gamma, Weibull, beta, and log normal distributions. Other features that one might consider removing from a one-semester curriculum include maximum likelihood estimation, prediction, and/or tolerance limits in Chapter 9. A one-semester curriculum has built-in flexibility, depending on the relative interest of the instructor in regression, analysis of variance, ex perimental design, and response surface methods (Chapter 15). There are several

discrete and continuous distributions (Chapters 5 and 6) that have applications in a variety of engineering and scientific areas.

Chapters 11 through 18 contain substantial material that can be added for the second semester of a two-semester course. The material on simple and multiple linear regression is in Chapters 11 and 12, respectively. Chapter 12 alone offers a substantial amount of flexibility. Multiple linear regression includes such “special topics” as categorical or indicator variables, sequential methods of model selection such as stepwise regression, the study of residuals for the detection of violations of assumptions, cross validation and the use of the PRESS statistic as well as C_p , and logistic regression. The use of orthogonal regressors, a precursor to the experimental design in Chapter 15, is highlighted. Chapters 13 and 14 offer a relatively large amount of material on analysis of variance (ANOVA) with fixed, random, and mixed models. Chapter 15 highlights the application of two-level designs in the context of full and fractional factorial experiments (2^k). Special screening designs are illustrated. Chapter 15 also features a new section on response surface methodology (RSM) to illustrate the use of experimental design for finding optimal process conditions. The fitting of a second order model through the use of a central composite design is discussed. RSM is expanded to cover the analysis of robust parameter design type problems. Noise variables are used to accommodate dual response surface models. Chapters 16, 17, and 18 contain a moderate amount of material on nonparametric statistics, quality control, and Bayesian inference.

Chapter 1 is an overview of statistical inference presented on a mathematically simple level. It has been expanded from the eighth edition to more thoroughly cover single-number statistics and graphical techniques. It is designed to give students a preliminary presentation of elementary concepts that will allow them to understand more involved details that follow. Elementary concepts in sampling, data collection, and experimental design are presented, and rudimentary aspects of graphical tools are introduced, as well as a sense of what is garnered from a data set. Stem-and-leaf plots and box-and-whisker plots have been added. Graphs are better organized and labeled. The discussion of uncertainty and variation in a system is thorough and well illustrated. There are examples of how to sort out the important characteristics of a scientific process or system, and these ideas are illustrated in practical settings such as manufacturing processes, biomedical studies, and studies of biological and other scientific systems. A contrast is made between the use of discrete and continuous data. Emphasis is placed on the use of models and the information concerning statistical models that can be obtained from graphical tools.

Chapters 2, 3, and 4 deal with basic probability as well as discrete and continuous random variables. Chapters 5 and 6 focus on specific discrete and continuous distributions as well as relationships among them. These chapters also highlight examples of applications of the distributions in real-life scientific and engineering studies. Examples, case studies, and a large number of exercises edify the student concerning the use of these distributions. Projects bring the practical use of these distributions to life through group work. Chapter 7 is the most theoretical chap ter in the text. It deals with transformation of random variables and will likely not be used unless the instructor wishes to teach a relatively theoretical course. Chapter 8 contains graphical material, expanding on the more elementary set of graphical tools presented and illustrated in Chapter 1. Probability plotting is discussed and illustrated with examples. The very important concept of sampling distributions is presented thoroughly, and illustrations are given that involve the central limit theorem and the distribution of a sample variance under normal, independent (i.i.d.) sampling. The *t* and *F* distributions are introduced to motivate their use in chapters to follow. New material in Chapter 8 helps the student to visualize the importance of hypothesis testing, motivating the concept of a *P*-value.

Chapter 9 contains material on one- and two-sample point and interval estimation. A thorough discussion with examples points out the contrast between the different types of intervals—confidence intervals, prediction intervals, and toler ance intervals. A case study illustrates the three types of statistical intervals in the context of a manufacturing situation. This case study highlights the differences among the intervals, their sources, and the assumptions made in their develop ment, as well as what type of scientific study or question requires the use of each one. A new approximation method has been added for the inference concerning a proportion. Chapter 10 begins with a basic presentation on the pragmatic mean ing of hypothesis testing, with emphasis on such fundamental concepts as null and alternative hypotheses, the role of probability and the *P*-value, and the power of a test. Following this, illustrations are given of tests concerning one and two sam ples under standard conditions. The two-sample *t*-test with paired observations is also described. A case study helps the student to develop a clear picture of what interaction among factors really means as well as the dangers that can arise when interaction between treatments and experimental units exists. At the end of Chapter 10 is a very important section that relates Chapters 9 and 10 (estimation and hypothesis testing) to Chapters 11 through 16, where statistical modeling is prominent. It is important that the student be aware of the strong connection.

Chapters 11 and 12 contain material on simple and multiple linear regression, respectively. Considerably more attention is given in this edition to the effect that collinearity among the regression variables plays. A situation is presented that shows how the role of a single regression variable can depend in large part on what regressors are in the model with it. The sequential model selection procedures (for ward, backward, stepwise, etc.) are then revisited in regard to this concept, and the rationale for using certain *P*-values with these procedures is provided. Chapter 12 offers material on nonlinear modeling with a special presentation of logistic regression, which has applications in engineering and the biological sciences. The material on multiple regression is quite extensive and thus provides considerable flexibility for the instructor, as indicated earlier. At the end of Chapter 12 is commentary relating that chapter to Chapters 14 and 15. Several features were added that provide a better understanding of the material in general. For example, the end-of-chapter material deals with cautions and difficulties one might encounter. It is pointed out that there are types of responses that occur naturally in practice (e.g. proportion responses, count responses, and several others) with which standard least squares regression should not be used because standard assumptions do not hold and violation of assumptions may induce serious errors. The suggestion is made that data transformation on the response may alleviate the problem in some cases. Flexibility is again available in Chapters 13 and 14, on the topic of analysis of variance. Chapter 13 covers one-factor ANOVA in the context of a completely randomized design. Complementary topics include tests on variances and multiple comparisons. Comparisons of treatments in blocks are highlighted, along with the topic of randomized complete blocks. Graphical methods are extended to ANOVA to aid the student in supplementing the formal inference with a pictorial type of inference that can aid scientists and engineers in presenting material. A new project is given in which students incorporate the appropriate randomization into each plan and use graphical techniques and *P*-values in reporting the results. Chapter 14 extends the material in Chapter 13 to accommodate two or more factors that are in a factorial structure. The ANOVA presentation in Chapter 14 includes work in both random and fixed effects models. Chapter 15 offers material associated with 2^k factorial designs; examples and case studies present the use of screening designs and special higher fractions of the 2^k . Two new and special features are the presentations of response surface methodology (RSM) and robust parameter design. These topics are linked in a case study that describes and illustrates a dual response surface design and analysis featuring the use of process mean and variance response surfaces.

Computer Software

Case studies, beginning in Chapter 8, feature computer printout and graphical material generated using both SAS and MINITAB. The inclusion of the computer reflects our belief that students should have the experience of reading and interpreting computer printout and graphics, even if the software in the text is not that which is used by the instructor. Exposure to more than one type of software can broaden the experience base for the student. There is no reason to believe that the software used in the course will be that which the student will be called upon to use in practice following graduation. Examples and case studies in the text are supplemented, where appropriate, by various types of residual plots, quantile plots, normal probability plots, and other plots. Such plots are particularly prevalent in Chapters 11 through 15.

Acknowledgments

We are indebted to those colleagues who reviewed the previous editions of this book and provided many helpful suggestions for this edition. They are David Groggel, *Miami University*; Lance Hemlow, *Raritan Valley Community College*; Ying Ji, *University of Texas at San Antonio*; Thomas Kline, *University of Northern Iowa*; Sheila Lawrence, *Rutgers University*; Luis Moreno, *Broome County*

Community College; Donald Waldman, *University of Colorado—Boulder*; and Marlene Will, *Spalding University*. We would also like to thank Delray Schulz, *Millersville Uni versity*; Roxane Burrows, *Hocking College*; and Frank Chmely for ensuring the accuracy of this text.

We would like to thank the editorial and production services provided by numerous people from Pearson, especially editor in chief Deirdre Lynch, acquisitions editor Patrick Barbera, Project Manager Christine Whitlock, Editorial Assistant Justin Billing, and copyeditor Sally Lifland. Many useful comments and sugges tions by proofreader Gail Magin are greatly appreciated. We thank the Virginia Tech Statistical Consulting Center, which was the source of many real-life data sets.

R.H.M.
S.L.M.
K.Y.

Acknowledgments for the Global Edition

Pearson would like to thank and acknowledge Neelesh Upadhye, *Indian Institute of Technology Madras*, Aneesh Kumar K., *Mahatma Gandhi College*, and Bindu P. P., *Government Arts and Science College*, for contributing to the Global Edition, and Abhishek K. Umrawal, *University of Delhi*, Olivia T.K. Choi, *The University of Hong Kong*, Mani Sankar, *East Point College of Engineering and Technology*, and Shalabh, *Indian Institute of Technology Kanpur*, for reviewing the Global Edition.

- The Adaptive Study Plan acts as a personal tutor, updating in real time based on student performance to provide personalized recommendations on what to work on next. With the new Companion Study Plan assignments, instructors can now assign the Study Plan as a prerequisite to a test or quiz, helping to guide students through concepts they need to master.

- Personalized Homework allows instructors to create homework assignments tailored to each student's specific needs, focused on just the topics they have not yet mastered.

Used by nearly 4 million students each year, the MyStatLab and MyMathLab family of products delivers consistent, measurable gains in student learning outcomes, retention, and subsequent course success.

Resources for Success

Instructor's Solutions Manual

The Instructor's Solutions Manual contains worked out solutions to all text exercises and is available for download from Pearson Education's Instructor's Resource Center **(www.pearsonglobaleditions.com/walpole)** and in MyStatLab.

PowerPoint Slides

The PowerPoint slides include most of the figures and tables from the text. Slides are available to download from Pearson Education's Instructor Resource Center **(www.pearsonglobaleditions.com/walpole)** and in MyStatLab.

MyStatLab™ Online Course (access code required)

MyStatLab from Pearson is the world's leading online resource for teaching and learning statistics; it integrates interactive homework, assessment, and media in a flexible, easy to use format. MyStatLab is a course management system that helps individual students succeed. It provides engaging experiences that personalize, stimulate, and measure learning for each student. Tools are embedded to make it easy to integrate statistical software into the course. And, it comes from an experienced partner with educational expertise and an eye on the future.

MyStatLab leverages the power of the web-based statistical software, StatCrunch™, and includes access to **www.StatCrunch.com.** To learn more about how MyStatLab combines proven learning applications with powerful assessment, visit **www.mystatlab.com** or contact your Pearson representative.

Chapter 1

Introduction to Statistics and Data Analysis

1.1 Overview: Statistical Inference, Samples, Populations, and the Role of Probability

Beginning in the 1980s and continuing into the 21st century, an inordinate amount of attention has been focused on *improvement of quality* in American industry. Much has been said and written about the Japanese "industrial miracle," which began in the middle of the 20th century. The Japanese were able to succeed where we and other countries had failed–namely, to create an atmosphere that allows the production of high-quality products. Much of the success of the Japanese has been attributed to the use of *statistical methods* and statistical thinking among management personnel.

Use of Scientific Data

The use of statistical methods in manufacturing, development of food products, computer software, energy sources, pharmaceuticals, and many other areas involves the gathering of information or scientific data. Of course, the gathering of data is nothing new. It has been done for well over a thousand years. Data have been collected, summarized, reported, and stored for perusal. However, there is a profound distinction between collection of scientific information and inferential statistics. It is the latter that has received rightful attention in recent decades.

The offspring of inferential statistics has been a large “toolbox” of statistical methods employed by statistical practitioners. These statistical methods are designed to contribute to the process of making scientific judgments in the face of uncertainty and variation. The product density of a particular material from a manufacturing process will not always be the same. Indeed, if the process involved is a batch process rather than continuous, there will be not only variation in material density among the batches that come off the line (batch-to-batch variation), but also within-batch variation. Statistical methods are used to analyze data from a process such as this one in order to gain more sense of where in the process changes may be made to improve the quality of the process. In this process, quality may well be defined in relation to closeness to a target density value in harmony with *what portion of the time* this closeness criterion is met. An engineer may be concerned with a specific instrument that is used to measure sulfur monoxide in the air during pollution studies. If the engineer has doubts about the effectiveness of the instrument, there are two sources of variation that must be dealt with. The first is the variation in sulfur monoxide values that are found at the same locale on the same day. The second is the variation between values observed and the true amount of sulfur monoxide that is in the air at the time. If either of these two sources of variation is exceedingly large (according to some standard set by the engineer), the instrument may need to be replaced. In a biomedical study of a new drug that reduces hypertension, 85% of patients experienced relief, while it is generally recognized that the current drug, or “old” drug, brings relief to 80% of patients that have chronic hypertension. However, the new drug is more expensive to make and may result in certain side effects. Should the new drug be adopted? This is a problem that is encountered (often with much more complexity) frequently by pharmaceutical firms in conjunction with the FDA (Federal Drug Administration). Again, the consideration of variation needs to be taken into account. The “85%” value is based on a certain number of patients chosen for the study. Perhaps if the study were repeated with new patients the observed number of “successes” would be 75%! It is the natural variation from study to study that must be taken into account in the decision process. Clearly this variation is important, since variation from patient to patient is endemic to the problem.

Variability in Scientific Data

In the problems discussed above the statistical methods used involve dealing with variability, and in each case the variability to be studied is that encountered in scientific data. If the observed product density in the process were always the same and were always on target, there would be no need for statistical methods. If the device for measuring sulfur monoxide always gives the same value and the value is accurate (i.e., it is correct), no statistical analysis is needed. If there were no patient-to-patient variability inherent in the response to the drug (i.e., it either always brings relief or not), life would be simple for scientists in the pharmaceutical firms and FDA and no statistician would be needed in the decision process. Statistics researchers have produced an enormous number of analytical methods that allow for analysis of data from systems like those described above. This reflects the true nature of the science that we call inferential statistics, namely, using techniques that allow us to go beyond merely reporting data to drawing conclusions (or inferences) about the scientific system. Statisticians make use of fundamental laws of probability and statistical inference to draw conclusions about scientific systems. Information is gathered in the form of samples, or collections of observations. The process of sampling is introduced in Chapter 2, and the discussion continues throughout the entire book.

Samples are collected from populations, which are collections of all individuals or individual items of a particular type. At times a population signifies a scientific system. For example, a manufacturer of computer boards may wish to eliminate defects. A sampling process may involve collecting information on 50 computer boards sampled randomly from the process. Here, the population is all computer boards manufactured by the firm over a specific period of time. If an improvement is made in the computer board process and a second sample of boards is collected, any conclusions drawn regarding the effectiveness of the change in pro cess should extend to the entire population of computer boards produced under the "improved process." In a drug experiment, a sample of patients is taken and each is given a specific drug to reduce blood pressure. The interest is focused on drawing conclusions about the population of those who suffer from hypertension.

Often, it is very important to collect scientific data in a systematic way, with planning being high on the agenda. At times the planning is, by necessity, quite limited. We often focus only on certain properties or characteristics of the items or objects in the population. Each characteristic has particular engineering or, say, biological importance to the "customer," the scientist or engineer who seeks to learn about the population. For example, in one of the illustrations above the quality of the process had to do with the product density of the output of a process. An engineer may need to study the effect of process conditions, temperature, humidity, amount of a particular ingredient, and so on. He or she can systematically move these factors to whatever levels are suggested according to whatever prescription or experimental design is desired. However, a forest scientist who is interested in a study of factors that influence wood density in a certain kind of tree cannot necessarily design an experiment. This case may require an observational study in which data are collected in the field but factor levels can not be preselected. Both of these types of studies lend themselves to methods of statistical inference. In the former, the quality of the inferences will depend on proper planning of the experiment. In the latter, the scientist is at the mercy of what can be gathered. For example, it is sad if an agronomist is interested in studying the effect of rainfall on plant yield and the data are gathered during a drought.

The importance of statistical thinking by managers and the use of statistical inference by scientific personnel is widely acknowledged. Research scientists gain much from scientific data. Data provide understanding of scientific phenomena. Product and process engineers learn a great deal in their off-line efforts to improve the process. They also gain valuable insight by gathering production data (on line monitoring) on a regular basis. This allows them to determine necessary modifications in order to keep the process at a desired level of quality.

There are times when a scientific practitioner wishes only to gain some sort of summary of a set of data represented in the sample. In other words, inferential statistics is not required. Rather, a set of single-number statistics or descriptive statistics is helpful. These numbers give a sense of center of the location of the data, variability in the data, and the general nature of the distribution of observations in the sample. Though no specific statistical methods leading to statistical inference are incorporated, much can be learned. At times, descriptive statistics are accompanied by graphics. Modern statistical software packages allow for computation of means, medians, standard deviations, and other single number statistics as well as production of graphs that show a “footprint” of the nature of the sample. Definitions and illustrations of the single-number statistics and graphs, including histograms, stem-and-leaf plots, scatter plots, dot plots, and box plots, will be given in sections that follow.

The Role of Probability

In this book, Chapters 2 to 6 deal with fundamental notions of probability. A thorough grounding in these concepts allows the reader to have a better understanding of statistical inference. Without some formalism of probability theory, the student cannot appreciate the true interpretation from data analysis through modern statistical methods. It is quite natural to study probability prior to studying statistical inference. Elements of probability allow us to quantify the strength or “confidence” in our conclusions. In this sense, concepts in probability form a major component that supplements statistical methods and helps us gauge the strength of the statistical inference. The discipline of probability, then, provides the transition between descriptive statistics and inferential methods. Elements of probability allow the conclusion to be put into the language that the science or engineering practitioners require. An example follows that will enable the reader to understand the notion of a *P*-value, which often provides the “bottom line” in the interpretation of results from the use of statistical methods.

Example 1.1: Suppose that an engineer encounters data from a manufacturing process in which 100 items are sampled and 10 are found to be defective. It is expected and anticipated that occasionally there will be defective items. Obviously these 100 items represent the sample. However, it has been determined that in the long run, the company can only tolerate 5% defective in the process. Now, the elements of prob ability allow the engineer to determine how conclusive the sample information is regarding the nature of the process. In this case, the population conceptually represents all possible items from the process. Suppose we learn that *if the process is acceptable*, that is, if it does produce items no more than 5% of which are de fective, there is a probability of 0.0282 of obtaining 10 or more defective items in a random sample of 100 items from the process. This small probability suggests that the process does, indeed, have a long-run rate of defective items that exceeds 5%. In other words, under the condition of an acceptable process, the sample in formation obtained would rarely occur. However, it did occur! Clearly, though, it would occur with a much higher probability if the process defective rate exceeded 5% by a significant amount.

From this example it becomes clear that the elements of probability aid in the translation of sample information into something conclusive or inconclusive about the scientific system. In fact, what was learned likely is alarming information to the engineer or manager. Statistical methods, which we will actually detail in Chapter 10, produced a *P*-value of 0.0282. The result suggests that the process very likely is not acceptable. The concept of a *P*-value is dealt with at length in succeeding chapters. The example that follows provides a second illustration.

Example 1.2: Often the nature of the scientific study will dictate the role that probability and deductive reasoning play in statistical inference. Exercise 9.40 on page 314 provides data associated with a study conducted at the Virginia Polytechnic Institute and State University on the development of a relationship between the roots of trees and the action of a fungus. Minerals are transferred from the fungus to the trees and sugars from the trees to the fungus. Two samples of 10 northern red oak seedlings were planted in a greenhouse, one containing seedlings treated with nitrogen and the other containing seedlings with no nitrogen. All other environmental conditions were held constant. All seedlings contained the fungus *Pisolithus tinctorus*. More details are supplied in Chapter 9. The stem weights in grams were recorded after the end of 140 days. The data are given in Table 1.1.

Table 1.1: Data Set for Example 1.2

No Nitrogen	Nitrogen
0.32	0.26
0.53	0.43
0.28	0.47
0.37	0.49
0.47	0.52
0.43	0.75
0.36	0.79
0.42	0.86
0.38	0.62
0.43	0.46

0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 Figure 1.1: A dot plot of stem weight data.

In this example there are two samples from two separate populations. The purpose of the experiment is to determine if the use of nitrogen has an influence on the growth of the roots. The study is a comparative study (i.e., we seek to compare the two populations with regard to a certain important characteristic). It is instructive to plot the data as shown in the dot plot of Figure 1.1. The ◦ values represent the “nitrogen” data and the × values represent the “no-nitrogen” data.

Notice that the general appearance of the data might suggest to the reader that, on average, the use of nitrogen increases the stem weight. Four nitrogen ob servations are considerably larger than any of the no-nitrogen observations. Most of the no-nitrogen observations appear to be below the center of the data. The appearance of the data set would seem to indicate that nitrogen is effective. But how can this be quantified? How can all of the apparent visual evidence be summa rized in some sense? As in the preceding example, the fundamentals of probability can be used. The conclusions may be summarized in a probability statement or *P*-value. We will not show here the statistical inference that produces the summary probability. As in Example 1.1, these methods will be discussed in Chapter 10. The issue revolves around the “probability that data like these could be observed” *given that nitrogen has no effect*, in other words, given that both samples were generated from the same population. Suppose that this probability is small, say 0.03. That would certainly be strong evidence that the use of nitrogen does indeed influence (apparently increases) average stem weight of the red oak seedlings.

How Do Probability and Statistical Inference Work Together?

It is important for the reader to understand the clear distinction between the discipline of probability, a science in its own right, and the discipline of inferen tial statistics. As we have already indicated, the use or application of concepts in probability allows real-life interpretation of the results of statistical inference. As a result, it can be said that statistical inference makes use of concepts in probability. One can glean from the two examples above that the sample information is made available to the analyst and, with the aid of statistical methods and elements of probability, conclusions are drawn about some feature of the population (the pro cess does not appear to be acceptable in Example 1.1, and nitrogen does appear to influence average stem weights in Example 1.2). Thus for a statistical problem, the sample along with inferential statistics allows us to draw conclu sions about the population, with inferential statistics making clear use of elements of probability. This reasoning is *inductive* in nature. Now as we move into Chapter 2 and beyond, the reader will note that, unlike what we do in our two examples here, we will not focus on solving statistical problems. Many examples will be given in which no sample is involved. There will be a population clearly described with all features of the population known. Then questions of im portance will focus on the nature of data that might hypothetically be drawn from the population. Thus, one can say that elements in probability allow us to draw conclusions about characteristics of hypothetical data taken from the population, based on known features of the population. This type of reasoning is *deductive* in nature. Figure 1.2 shows the fundamental relationship between probability and inferential statistics.

Probability

Population Sample

Statistical Inference

Figure 1.2: Fundamental relationship between probability and inferential statistics.

Now, in the grand scheme of things, which is more important, the field of probability or the field of statistics? They are both very important and clearly are complementary. The only certainty concerning the pedagogy of the two disciplines lies in the fact that if statistics is to be taught at more than merely a “cookbook” level, then the discipline of probability must be taught first. This rule stems from the fact that nothing can be learned about a population from a sample until the analyst learns the rudiments of uncertainty in that sample. For example, consider Example 1.1. The question centers around whether or not the population, defined by the process, is no more than 5% defective. In other words, the conjecture is that on the average 5 out of 100 items are defective. Now, the sample contains 100 items and 10 are defective. Does this support the conjecture or refute it? On the surface it would appear to be a refutation of the conjecture because 10 out of 100 seem to be “a bit much.” But without elements of probability, how do we know? Only through the study of material in future chapters will we learn the conditions under which the process is acceptable (5% defective). The probability of obtaining 10 or more defective items in a sample of 100 is 0.0282.

We have given two examples where the elements of probability provide a sum mary that the scientist or engineer can use as evidence on which to build a decision. The bridge between the data and the conclusion is, of course, based on foundations of statistical inference, distribution theory, and sampling distributions discussed in future chapters.

1.2 Sampling Procedures; Collection of Data

In Section 1.1 we discussed very briefly the notion of sampling and the sampling process. While sampling appears to be a simple concept, the complexity of the questions that must be answered about the population or populations necessitates that the sampling process be very complex at times. While the notion of sampling is discussed in a technical way in Chapter 8, we shall endeavor here to give some common-sense notions of sampling. This is a natural transition to a discussion of the concept of variability.

Simple Random Sampling

The importance of proper sampling revolves around the degree of confidence with which the analyst is able to answer the questions being asked. Let us assume that only a single population exists in the problem. Recall that in Example 1.2 two populations were involved. Simple random sampling implies that any particular sample of a specified *sample size* has the same chance of being selected as any other sample of the same size. The term sample size simply means the number of elements in the sample. Obviously, a table of random numbers can be utilized in sample selection in many instances. The virtue of simple random sampling is that it aids in the elimination of the problem of having the sample reflect a different (possibly more confined) population than the one about which inferences need to be made. For example, a sample is to be chosen to answer certain questions regarding political preferences in a certain state in the United States. The sample involves the choice of, say, 1000 families, and a survey is to be conducted. Now, suppose it turns out that random sampling is not used. Rather, all or nearly all of the 1000 families chosen live in an urban setting. It is believed that political preferences in rural areas differ from those in urban areas. In other words, the sample drawn actually confined the population and thus the inferences need to be confined to the “limited population,” and in this case confining may be undesirable. If, indeed, the inferences need to be made about the state as a whole, the sample of size 1000 described here is often referred to as a biased sample.

As we hinted earlier, simple random sampling is not always appropriate. Which alternative approach is used depends on the complexity of the problem. Often, for example, the sampling units are not homogeneous and naturally divide themselves into nonoverlapping groups that are homogeneous. These groups are called *strata*, and a procedure called *stratified random sampling* involves random selection of a sample *within* each stratum. The purpose is to be sure that each of the strata is neither over- nor underrepresented. For example, suppose a sample survey is conducted in order to gather preliminary opinions regarding a bond referendum that is being considered in a certain city. The city is subdivided into several ethnic groups which represent natural strata. In order not to disregard or overrepresent any group, separate random samples of families could be chosen from each group.

Experimental Design

The concept of randomness or random assignment plays a huge role in the area of experimental design, which was introduced very briefly in Section 1.1 and is an important staple in almost any area of engineering or experimental science. This will be discussed at length in Chapters 13 through 15. However, it is instructive to give a brief presentation here in the context of random sampling. A set of so-called treatments or treatment combinations becomes the populations to be studied or compared in some sense. An example is the nitrogen versus no-nitrogen treat ments in Example 1.2. Another simple example would be "placebo" versus "active drug," or in a corrosion fatigue study we might have treatment combinations that involve specimens that are coated or uncoated as well as conditions of low or high humidity to which the specimens are exposed. In fact, there are four treatment or factor combinations (i.e., 4 populations), and many scientific questions may be asked and answered through statistical and inferential methods. Consider first the situation in Example 1.2. There are 20 diseased seedlings involved in the exper iment. It is easy to see from the data themselves that the seedlings are different from each other. Within the nitrogen group (or the no-nitrogen group) there is considerable variability in the stem weights. This variability is due to what is generally called the experimental unit. This is a very important concept in in ferential statistics, in fact one whose description will not end in this chapter. The nature of the variability is very important. If it is too large, stemming from a condition of excessive nonhomogeneity in experimental units, the variability will "wash out" any detectable difference between the two populations. Recall that in this case that did not occur.

The dot plot in Figure 1.1 and *P*-value indicated a clear distinction between these two conditions. What role do those experimental units play in the data taking process itself? The common-sense and, indeed, quite standard approach is to assign the 20 seedlings or experimental units randomly to the two treat ments or conditions. In the drug study, we may decide to use a total of 200 available patients, patients that clearly will be different in some sense. They are the experimental units. However, they all may have the same chronic condition for which the drug is a potential treatment. Then in a so-called completely ran domized design, 100 patients are assigned randomly to the placebo and 100 to the active drug. Again, it is these experimental units within a group or treatment that produce the variability in data results (i.e., variability in the measured result), say blood pressure, or whatever drug efficacy value is important. In the corrosion fatigue study, the experimental units are the specimens that are the subjects of the corrosion.

Why Assign Experimental Units Randomly?

What is the possible negative impact of not randomly assigning experimental units to the treatments or treatment combinations? This is seen most clearly in the case of the drug study. Among the characteristics of the patients that produce variability in the results are age, gender, and weight. Suppose merely by chance the placebo group contains a sample of people that are predominately heavier than those in the treatment group. Perhaps heavier individuals have a tendency to have a higher blood pressure. This clearly biases the result, and indeed, any result obtained through the application of statistical inference may have little to do with the drug and more to do with differences in weights among the two samples of patients.

We should emphasize the attachment of importance to the term variability. Excessive variability among experimental units "camouflages" scientific findings. In future sections, we attempt to characterize and quantify measures of variability. In sections that follow, we introduce and discuss specific quantities that can be computed in samples; the quantities give a sense of the nature of the sample with respect to center of location of the data and variability in the data. A discussion of several of these single-number measures serves to provide a preview of what statistical information will be important components of the statistical methods that are used in future chapters. These measures that help characterize the nature of the data set fall into the category of descriptive statistics. This material is a prelude to a brief presentation of pictorial and graphical methods that go even further in characterization of the data set. The reader should understand that the statistical methods illustrated here will be used throughout the text. In order to offer the reader a clearer picture of what is involved in experimental design studies, we offer Example 1.3.

Example 1.3: A corrosion study was made in order to determine whether coating an aluminum metal with a corrosion retardation substance reduced the amount of corrosion. The coating is a protectant that is advertised to minimize fatigue damage in this type of material. Also of interest is the influence of humidity on the amount of corrosion. A corrosion measurement can be expressed in thousands of cycles to failure. Two levels of coating, no coating and chemical corrosion coating, were used. In addition, the two relative humidity levels are 20% relative humidity and 80% relative humidity.

The experiment involves four treatment combinations that are listed in the table that follows. There are eight experimental units used, and they are aluminum specimens prepared; two are assigned randomly to each of the four treatment combinations. The data are presented in Table 1.2.

The corrosion data are averages of two specimens. A plot of the averages is pictured in Figure 1.3. A relatively large value of cycles to failure represents a small amount of corrosion. As one might expect, an increase in humidity appears to make the corrosion worse. The use of the chemical corrosion coating procedure appears to reduce corrosion.

In this experimental design illustration, the engineer has systematically selected the four treatment combinations. In order to connect this situation to concepts with which the reader has been exposed to this point, it should be assumed that the

Table 1.2: Data for Example 1.3

Coating	Humidity	Average Corrosion in Thousands of Cycles to Failure
Uncoated	20%	975
	80%	350

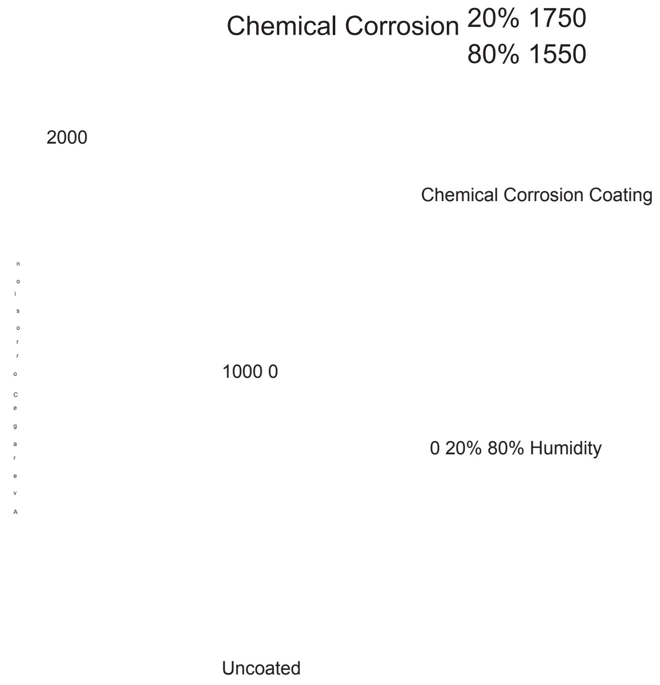

Figure 1.3: Corrosion results for Example 1.3.

conditions representing the four treatment combinations are four separate popula tions and that the two corrosion values observed for each population are important pieces of information. The importance of the average in capturing and summariz ing certain features in the population will be highlighted in Section 1.3. While we might draw conclusions about the role of humidity and the impact of coating the specimens from the figure, we cannot truly evaluate the results from an analyti cal point of view without taking into account the *variability around* the average. Again, as we indicated earlier, if the two corrosion values for each treatment com bination are close together, the picture in Figure 1.3 may be an accurate depiction. But if each corrosion value in the figure is an average of two values that are widely dispersed, then this variability may, indeed, truly "wash away" any information that appears to come through when one observes averages only. The foregoing example illustrates these concepts:

(1) random assignment of treatment combinations (coating, humidity) to experi mental units (specimens)

(2) the use of sample averages (average corrosion values) in summarizing sample information

(3) the need for consideration of measures of variability in the analysis of any sample or sets of samples

This example suggests the need for what follows in Sections 1.3 and 1.4, namely, descriptive statistics that indicate measures of center of location in a set of data, and those that measure variability.

1.3 Measures of Location: The Sample Mean and Median

Measures of location are designed to provide the analyst with some quantitative values of where the center, or some other location, of data is located. In Example 1.2, it appears as if the center of the nitrogen sample clearly exceeds that of the no-nitrogen sample. One obvious and very useful measure is the sample mean. The mean is simply a numerical average.

Definition 1.1: Definition 1.2:

Suppose that the observations
mean, denoted by $\bar{x}$, is

$\bar{x}$ = n

x_i

i=1

There are other measures of central tendency that are discussed in detail in future chapters. One important measure is the sample median. The purpose of the sample median is to reflect the central tendency of the sample in such a way that it is uninfluenced by extreme values or outliers.

Given that the observations in
increasing order of magnitude,

$x_($

$\tilde{x}$ =

$\frac{1}{2}$ $(x_{n}$

As an example, suppose the data set is the following: 1.7, 2.2, 3.9, 3.11, and 14.7. The sample mean and median are, respectively,

$\bar{x}$ = 5.12, $\tilde{x}$ = 3.11.

Clearly, the mean is influenced considerably by the presence of the extreme obser vation, 14.7, whereas the median places emphasis on the true “center” of the data set. In the case of the two-sample data set of Example 1.2, the two measures of central tendency for the individual samples are

$\bar{x}$ (no nitrogen) = 0.399 gram,

$\tilde{x}$ (no nitrogen) = 0.38 + 0.42

2 = 0.400 gram,

$\bar{x}$ (nitrogen) = 0.565 gram,

$\tilde{x}$ (nitrogen) = 0.49 + 0.52

2 = 0.505 gram.

Clearly there is a difference in concept between the mean and median. It may be of interest to the reader with an engineering background that the sample mean

is the centroid of the data in a sample. In a sense, it is the point at which a fulcrum can be placed to balance a system of "weights" which are the locations of the individual data. This is shown in Figure 1.4 with regard to the with-nitrogen sample.

x- 0.565

0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 Figure 1.4: Sample

mean as a centroid of the with-nitrogen stem weight.

In future chapters, the basis for the computation of $\bar{x}$ is that of an estimate of the population mean. As we indicated earlier, the purpose of statistical inference is to draw conclusions about population characteristics or parameters and estimation is a very important feature of statistical inference.

The median and mean can be quite different from each other. Note, however, that in the case of the stem weight data the sample mean value for no-nitrogen is quite similar to the median value.

Other Measures of Locations

There are several other methods of quantifying the center of location of the data in the sample. We will not deal with them at this point. For the most part, alternatives to the sample mean are designed to produce values that represent compromises between the mean and the median. Rarely do we make use of these other measures. However, it is instructive to discuss one class of estimators, namely the class of trimmed means. A trimmed mean is computed by "trimming away" a certain percent of both the largest and the smallest set of values. For example, the 10% trimmed mean is found by eliminating the largest 10% and smallest 10% and computing the average of the remaining values. For example, in the case of the stem weight data, we would eliminate the largest and smallest since the sample size is 10 for each sample. So for the without-nitrogen group the 10% trimmed mean is given by

$$\bar{x}_{tr(10)} = \frac{0.32 + 0.37 + 0.47 + 0.43 + 0.36 + 0.42 + 0.38 + 0.43}{8} = 0.39750,$$

and for the 10% trimmed mean for the with-nitrogen group we have

$$\bar{x}_{tr(10)} = \frac{0.43 + 0.47 + 0.49 + 0.52 + 0.75 + 0.79 + 0.62 + 0.46}{8} = 0.56625.$$

Note that in this case, as expected, the trimmed means are close to both the mean and the median for the individual samples. The trimmed mean is, of course, more insensitive to outliers than the sample mean but not as insensitive as the median. On the other hand, the trimmed mean approach makes use of more information than the sample median. Note that the sample median is, indeed, a special case of the trimmed mean in which all of the sample data are eliminated apart from the middle one or two observations.

Exercises

1.1 The following measurements were recorded for the drying time, in hours, of a certain brand of latex paint.

3.4 2.5 4.8 2.9 3.6
2.8 3.3 5.6 3.7 2.8
4.4 4.0 5.2 3.0 4.8

Assume that the measurements are a simple random sample.

(a) What is the sample size for the above sample?

(b) Calculate the sample mean for these data. (c) Calculate the sample median.

(d) Plot the data by way of a dot plot.

(e) Compute the 20% trimmed mean for the above data set.

(f) Is the sample mean for these data more or less descriptive as a center of location than the trimmed mean?

1.2 According to the journal *Chemical Engineering*, an important property of a fiber is its water ab sorbency. A random sample of 20 pieces of cotton fiber was taken and the absorbency on each piece was mea sured. The following are the absorbency values:

18.71 21.41 20.72 21.81 19.29 22.43 20.17
23.71 19.44 20.50 18.92 20.33 23.00 22.85
19.25 21.77 22.11 19.77 18.04 21.12

(a) Calculate the sample mean and median for the above sample values.

(b) Compute the 10% trimmed mean.

(c) Do a dot plot of the absorbency data. (d) Using only the values of the mean, median, and trimmed mean, do you have evidence of outliers in the data?

1.3 A certain polymer is used for evacuation systems for aircraft. It is important that the polymer be re sistant to the aging process. Twenty specimens of the polymer were used in an experiment. Ten were as signed randomly to be exposed to an accelerated batch aging process that involved exposure to high temperatures for 10 days. Measurements of tensile strength of the specimens were made, and the following data were recorded on tensile strength in psi:

No aging: 227 222 218 217 225
218 216 229 228 221
Aging: 219 214 215 211 209
218 203 204 201 205

(a) Do a dot plot of the data.

(b) From your plot, does it appear as if the aging process has had an effect on the tensile strength of this polymer? Explain.

(c) Calculate the sample mean tensile strength of the two samples.

(d) Calculate the median for both. Discuss the similarity or lack of similarity between the mean and median of each group.

1.4 In a study conducted by the Department of Mechanical Engineering at Virginia Tech, the steel rods supplied by two different companies were compared. Ten sample springs were made out of the steel rods supplied by each company, and a measure of flexibility was recorded for each. The data are as follows:

Company A: 9.3 8.8 6.8 8.7 8.5 6.7 8.0
6.5 9.2 7.0
Company B: 11.0 9.8 9.9 10.2 10.1 9.7
11.0 11.1 10.2 9.6

(a) Calculate the sample mean and median for the data for the two companies.

(b) Plot the data for the two companies on the same line and give your impression regarding any apparent differences between the two companies.

1.5 Twenty adult males between the ages of 30 and 40 participated in a study to evaluate the effect of a specific health regimen involving diet and exercise on the blood cholesterol. Ten were randomly selected to be a control group, and ten others were assigned to take part in the regimen as the treatment group for a period of 6 months. The following data show the reduction in cholesterol experienced for the time period for the 20 subjects:

Control group: 7 3 −4 14 2 5 22 −7 95
Treatment group: −6 5 9 44 12 37 5 3 3

(a) Do a dot plot of the data for both groups on the same graph.

(b) Compute the mean, median, and 10% trimmed mean for both groups.

(c) Explain why the difference in means suggests one conclusion about the effect of the regimen, while the difference in medians or trimmed means suggests a different conclusion.

1.6 The tensile strength of silicone rubber is thought to be a function of curing temperature. A study was carried out in which samples of 12 specimens of the rubber were prepared using curing temperatures of 20°C and 45°C. The data below show the tensile strength values in megapascals.

20°C: 2.07 2.14 2.22 2.03 2.21 2.03 2.05 2.18
2.09 2.14 2.11 2.02
45°C: 2.52 2.15 2.49 2.03 2.37 2.05 1.99 2.42
2.08 2.42 2.29 2.01

(a) Show a dot plot of the data with both low and high temperature tensile strength values.

(b) Compute sample mean tensile strength for both samples.

(c) Does it appear as if curing temperature has an influence on tensile strength, based on the plot? Comment further.

(d) Does anything else appear to be influenced by an increase in curing temperature? Explain.

1.4 Measures of Variability

Sample variability plays an important role in data analysis. Process and product variability is a fact of life in engineering and scientific systems: The control or reduction of process variability is often a source of major difficulty. More and more process engineers and managers are learning that product quality and, as a result, profits derived from manufactured products are very much a function of process variability. As a result, much of Chapters 9 through 15 deals with data analysis and modeling procedures in which sample variability plays a major role. Even in small data analysis problems, the success of a particular statistical method may depend on the magnitude of the variability among the observations in the sample. Measures of location in a sample do not provide a proper summary of the nature of a data set. For instance, in Example 1.2 we cannot conclude that the use of nitrogen enhances growth without taking sample variability into account.

While the details of the analysis of this type of data set are deferred to Chapter 9, it should be clear from Figure 1.1 that variability among the no-nitrogen observations and variability among the nitrogen observations are certainly of some consequence. In fact, it appears that the variability within the nitrogen sample is larger than that of the no-nitrogen sample. Perhaps there is something about the inclusion of nitrogen that not only increases the stem height ($\bar{x}$ of 0.565 gram compared to an $\bar{x}$ of 0.399 gram for the no-nitrogen sample) but also increases the variability in stem height (i.e., renders the stem height more inconsistent).

As another example, contrast the two data sets below. Each contains two samples and the difference in the means is roughly the same for the two samples, but data set B seems to provide a much sharper contrast between the two populations from which the samples were taken. If the purpose of such an experiment is to detect differences between the two populations, the task is accomplished in the case of data set B. However, in data set A the large variability *within* the two samples creates difficulty. In fact, it is not clear that there is a distinction *between* the two populations.

Data set A: X X X X X X 0 X X 0 0 X X X 0 0 0 0 0 0 0 0

$\bar{x}_X$ $\bar{x}_0$

Data set B: X X X X X X X X X X X 0 0 0 0 0 0 0 0 0 0 0 $\bar{x}_X$ $\bar{x}_0$

Sample Range and Sample Standard Deviation

Just as there are many measures of central tendency or location, there are many measures of spread or variability. Perhaps the simplest one is the sample range $X_{max} - X_{min}$. The range can be very useful and is discussed at length in Chapter 17 on *statistical quality control*. The sample measure of spread that is used most often is the sample standard deviation. We again let $x_1, x_2, \ldots, x_n$ denote sample values.

Definition 1.3:

The sample variance, denoted s^2

The sample standard deviation s^2, that is,

It should be clear to the reader that the sample standard deviation is, in fact, a measure of variability. Large variability in a data set produces relatively large values of $(x - \bar{x})^2$ and thus a large sample variance. The quantity $n - 1$ is often called the degrees of freedom associated with the variance estimate. In this simple example, the degrees of freedom depict the number of independent pieces of information available for computing variability. For example, suppose that we wish to compute the sample variance and standard deviation of the data set (5, 17, 6, 4). The sample average is $\bar{x} = 8$. The computation of the variance involves

$$(5 - 8)^2 + (17 - 8)^2 + (6 - 8)^2 + (4 - 8)^2 = (-3)^2 + 9^2 + (-2)^2 + (-4)^2.$$

The quantities inside parentheses sum to zero. In general, $^{n}{}_{i=1}(x_i - \bar{x}) = 0$ (see Exercise 1.16 on page 51). Then the computation of a sample variance does not involve n independent squared deviations from the mean $\bar{x}$. In fact, since the last value of $x - \bar{x}$ is determined by the initial $n - 1$ of them, we say that these are $n - 1$ "pieces of information" that produce s^2. Thus, there are $n - 1$ degrees of freedom rather than n degrees of freedom for computing a sample variance.

Example 1.4: In an example discussed extensively in Chapter 10, an engineer is interested in testing the "bias" in a pH meter. Data are collected on the meter by measuring the pH of a neutral substance (pH = 7.0). A sample of size 10 is taken, with results given by

$$7.07\ 7.00\ 7.10\ 6.97\ 7.00\ 7.03\ 7.01\ 7.01\ 6.98\ 7.08.$$

The sample mean $\bar{x}$ is given by

$$\bar{x} = \frac{7.07 + 7.00 + 7.10 + \cdots + 7.08}{10} = 7.0250.$$

The sample variance s^2 is given by

$$s^2 = \frac{1}{9}[(7.07 - 7.025)^2 + (7.00 - 7.025)^2 + (7.10 - 7.025)^2 + \cdots + (7.08 - 7.025)^2] = 0.001939.$$

As a result, the sample standard deviation is given by

$$s = \sqrt{0.001939} = 0.044.$$

So the sample standard deviation is 0.0440 with $n - 1 = 9$ degrees of freedom.

Units for Standard Deviation and Variance

It should be apparent from Definition 1.3 that the variance is a measure of the average squared deviation from the mean $\bar{x}$. We use the term *average squared deviation* even though the definition makes use of a division by degrees of freedom $n - 1$ rather than n. Of course, if n is large, the difference in the denominator is inconsequential. As a result, the sample variance possesses units that are the square of the units in the observed data whereas the sample standard deviation is found in linear units. As an example, consider the data of Example 1.2. The stem weights are measured in grams. As a result, the sample standard deviations are in grams and the variances are measured in grams^2. In fact, the individual standard deviations are 0.0728 gram for the no-nitrogen case and 0.1867 gram for the nitrogen group. Note that the standard deviation does indicate considerably larger variability in the nitrogen sample. This condition was displayed in Figure 1.1.

Which Variability Measure Is More Important?

As we indicated earlier, the sample range has applications in the area of statistical quality control. It may appear to the reader that the use of both the sample variance and the sample standard deviation is redundant. Both measures reflect the same concept in measuring variability, but the sample standard deviation measures variability in linear units whereas the sample variance is measured in squared units. Both play huge roles in the use of statistical methods. Much of what is accomplished in the context of statistical inference involves drawing conclusions about characteristics of populations. Among these characteristics are constants which are called population parameters. Two important parameters are the population mean and the population variance. The sample variance plays an explicit role in the statistical methods used to draw inferences about the population variance. The sample standard deviation has an important role along with the sample mean in inferences that are made about the population mean. In general, the variance is considered more in inferential theory, while the standard deviation is used more in applications.

Exercises

1.7 Consider the drying time data for Exercise 1.1 on page 33. Compute the sample variance and sample standard deviation.

1.8 Compute the sample variance and standard deviation for the water absorbency data of Exercise 1.2 on page 33.

1.9 Exercise 1.3 on page 33 showed tensile strength data for two samples, one in which specimens were exposed to an aging process and one in which there was no aging of the specimens.

(a) Calculate the sample variance as well as standard deviation in tensile strength for both samples. (b) Does there appear to be any evidence that aging affects the variability in tensile strength? (See also the plot for Exercise 1.3 on page 33.)

1.10 For the data of Exercise 1.4 on page 33, compute both the mean and the variance in “flexibility” for both company A and company B. Does there ap pear to be a difference in flexibility between company A and company B?

1.11 Consider the data in Exercise 1.5 on page 33. Compute the sample variance and the sample standard deviation for both control and treatment groups.

1.12 For Exercise 1.6 on page 33, compute the sample standard deviation in tensile strength for the samples separately for the two temperatures. Does it appear as if an increase in temperature influences the variability in tensile strength? Explain.

1.5 Discrete and Continuous Data

Statistical inference through the analysis of observational studies or designed experiments is used in many scientific areas. The data gathered may be discrete or continuous, depending on the area of application. For example, a chemical engineer may be interested in conducting an experiment that will lead to conditions where yield is maximized. Here, of course, the yield may be in percent or grams/pound, measured on a continuum. On the other hand, a toxicologist conducting a combination drug experiment may encounter data that are binary in nature (i.e., the patient either responds or does not).

Great distinctions are made between discrete and continuous data in the prob ability theory that allow us to draw statistical inferences. Often applications of statistical inference are found when the data are *count data*. For example, an en gineer may be interested in studying the number of radioactive particles passing through a counter in, say, 1 millisecond. Personnel responsible for the efficiency of a port facility may be interested in the properties of the number of oil tankers arriving each day at a certain port city. In Chapter 5, several distinct scenarios, leading to different ways of handling data, are discussed for situations with count data.

Special attention even at this early stage of the textbook should be paid to some details associated with binary data. Applications requiring statistical analysis of binary data are voluminous. Often the measure that is used in the analysis is the *sample proportion*. Obviously the binary situation involves two categories. If there are n units involved in the data and x is defined as the number that fall into category 1, then $n - x$ fall into category 2. Thus, x/n is the sample proportion in category 1, and $1 - x/n$ is the sample proportion in category 2. In the biomedical application, 50 patients may represent the sample units, and if 20 out of 50 experienced an improvement in a stomach ailment (common to all 50) after all were given the drug, then $\frac{20}{50} = 0.4$ is the sample proportion for which

the drug was a success and $1 - 0.4 = 0.6$ is the sample proportion for which the drug was not successful. Actually the basic numerical measurement for binary data is generally denoted by either 0 or 1. For example, in our medical example, a successful result is denoted by a 1 and a nonsuccess a 0. As a result, the sample proportion is actually a sample mean of the ones and zeros. For the successful category,

$$50 = 1+1+0+\cdots+0+1$$

$$x_1 + x_2 + \cdots + x_{50}$$

$$50 = \frac{20}{50} = 0.4.$$

What Kinds of Problems Are Solved in Binary Data Situations?

The kinds of problems facing scientists and engineers dealing in binary data are not a great deal unlike those seen where continuous measurements are of interest. However, different techniques are used since the statistical properties of sample proportions are quite different from those of the sample means that result from averages taken from continuous populations. Consider the example data in Ex ercise 1.6 on page 33. The statistical problem underlying this illustration focuses on whether an intervention, say, an increase in curing temperature, will alter the population mean tensile strength associated with the silicone rubber process. On the other hand, in a quality control area, suppose an automobile tire manufacturer reports that a shipment of 5000 tires selected randomly from the process results in 100 of them showing blemishes. Here the sample proportion is $\frac{100}{5000} = 0.02$.

Following a change in the process designed to reduce blemishes, a second sample of 5000 is taken and 90 tires are blemished. The sample proportion has been reduced to $\frac{90}{5000} = 0.018$. The question arises, “Is the decrease in the sample proportion from 0.02 to 0.018 substantial enough to suggest a real improvement in the population proportion?” Both of these illustrations require the use of the statistical properties of sample averages—one from samples from a continuous population, and the other from samples from a discrete (binary) population. In both cases, the sample mean is an estimate of a population parameter, a population mean in the first illustration (i.e., mean tensile strength), and a population proportion in the second case (i.e., proportion of blemished tires in the population). So here we have sample estimates used to draw scientific conclusions regarding population parameters. As we indicated in Section 1.3, this is the general theme in many practical problems using statistical inference.

1.6 Statistical Modeling, Scientific Inspection, and Graphical Diagnostics

Often the end result of a statistical analysis is the estimation of parameters of a postulated model. This is natural for scientists and engineers since they often deal in modeling. A statistical model is not deterministic but, rather, must entail some probabilistic aspects. A model form is often the foundation of assumptions that are made by the analyst. For example, in Example 1.2 the scientist may wish to draw some level of distinction between the nitrogen and no-nitrogen populations through the sample information. The analysis may require a certain model for the data, for example, that the two samples come from normal or Gaussian distributions. See Chapter 6 for a discussion of the normal distribution.

Obviously, the user of statistical methods cannot generate sufficient information or experimental data to characterize the population totally. But sets of data are often used to learn about certain properties of the population. Scientists and engineers are accustomed to dealing with data sets. The importance of characterizing or *summarizing* the nature of collections of data should be obvious. Often a summary of a collection of data via a graphical display can provide insight regarding the system from which the data were taken. For instance, in Sections 1.1 and 1.3, we have shown dot plots.

In this section, the role of sampling and the display of data for enhancement of statistical inference is explored in detail. We merely introduce some simple but often effective displays that complement the study of statistical populations.

Scatter Plot

At times the model postulated may take on a somewhat complicated form. Consider, for example, a textile manufacturer who designs an experiment where cloth specimen that contain various percentages of cotton are produced. Consider the data in Table 1.3.

Table 1.3: Tensile Strength

Cotton Percentage	Tensile Strength
15	7, 7, 9, 8, 10
20	19, 20, 21, 20, 22
25	21, 21, 17, 19, 20
30	8, 7, 8, 9, 10

Five cloth specimens are manufactured for each of the four cotton percentages. In this case, both the model for the experiment and the type of analysis used should take into account the goal of the experiment and important input from the textile scientist. Some simple graphics can shed important light on the clear distinction between the samples. See Figure 1.5; the sample means and variability are depicted nicely in the scatter plot. One possible goal of this experiment is simply to determine which cotton percentages are truly distinct from the others. In other words, as in the case of the nitrogen/no-nitrogen data, for which cotton percentages are there clear distinctions between the populations or, more specifically, between the population means? In this case, perhaps a reasonable model is that each sample comes from a normal distribution. Here the goal is very much like that of the nitrogen/no-nitrogen data except that more samples are involved. The formalism of the analysis involves notions of hypothesis testing discussed in Chapter 10. Incidentally, this formality is perhaps not necessary in light of the diagnostic plot. But does this describe the real goal of the experiment and hence the proper approach to data analysis? It is likely that the scientist anticipates the existence of a *maximum population mean tensile strength* in the range of cot ton concentration in the experiment. Here the analysis of the data should revolve around a different type of model, one that postulates a type of structure relating the population mean tensile strength to the cotton concentration. In other words, a model may be written

$$\mu_{t,c} = \beta_0 + \beta_1 C + \beta_2 C^2 ,$$

where $\mu_{t,c}$ is the population mean tensile strength, which varies with the amount of cotton in the product C. The implication of this model is that for a fixed cotton level, there is a population of tensile strength measurements and the population mean is $\mu_{t,c}$. This type of model, called a regression model, is discussed in Chapters 11 and 12. The functional form is chosen by the scientist. At times the data analysis may suggest that the model be changed. Then the data analyst "entertains" a model that may be altered after some analysis is done. The use of an empirical model is accompanied by estimation theory, where β_0 , β_1 , and β_2 are estimated by the data. Further, statistical inference can then be used to determine model adequacy.

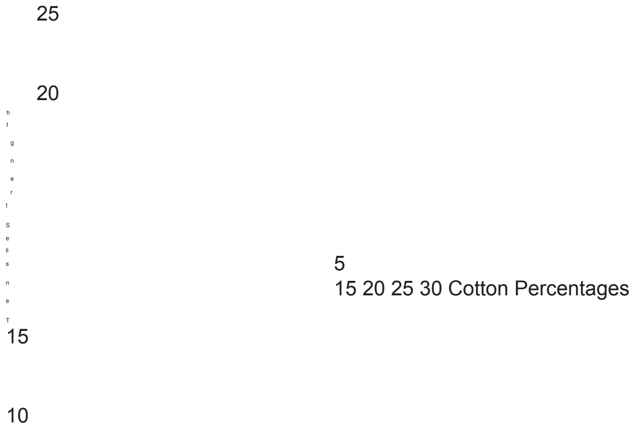

Figure 1.5: Scatter plot of tensile strength and cotton percentages.

Two points become evident from the two data illustrations here: (1) The type of model used to describe the data often depends on the goal of the experiment; and (2) the structure of the model should take advantage of nonstatistical scientific input. A selection of a model represents a fundamental assumption upon which the resulting statistical inference is based. It will become apparent throughout the book how important graphics can be. Often, plots can illustrate information that allows the results of the formal statistical inference to be better communicated to the scientist or engineer. At times, plots or exploratory data analysis can teach the analyst something not retrieved from the formal analysis. Almost any formal analysis requires assumptions that evolve from the model of the data. Graphics can nicely highlight violation of assumptions that would otherwise go unnoticed. Throughout the book, graphics are used extensively to supplement formal data analysis. The following sections reveal some graphical tools that are useful in exploratory or descriptive data analysis.

Stem-and-Leaf Plot

Statistical data, generated in large masses, can be very useful for studying the behavior of the distribution if presented in a combined tabular and graphic display called a stem-and-leaf plot.

To illustrate the construction of a stem-and-leaf plot, consider the data of Table 1.4, which specifies the "life" of 40 similar car batteries recorded to the nearest tenth of a year. The batteries are guaranteed to last 3 years. First, split each observation into two parts consisting of a stem and a leaf such that the stem represents the digit preceding the decimal and the leaf corresponds to the decimal part of the number. In other words, for the number 3.7, the digit 3 is designated the stem and the digit 7 is the leaf. The four stems 1, 2, 3, and 4 for our data are listed vertically on the left side in Table 1.5; the leaves are recorded on the right side opposite the appropriate stem value. Thus, the leaf 6 of the number 1.6 is recorded opposite the stem 1; the leaf 5 of the number 2.5 is recorded opposite the stem 2; and so forth. The number of leaves recorded opposite each stem is summarized under the frequency column.

Table 1.4: Car Battery Life

2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6
3.4	1.6	3.1	3.3	3.8	3.1	4.7	3.7
2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1
3.3	3.1	3.7	4.4	3.2	4.1	1.9	3.4
4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5

Table 1.5: Stem-and-Leaf Plot of Battery Life

	Stem Leaf Frequency
1	7778899 4
69	11234577
2	2
25669	5
3	25 8
001111222333444556	

The stem-and-leaf plot of Table 1.5 contains only four stems and consequently does not provide an adequate picture of the distribution. To remedy this problem, we need to increase the number of stems in our plot. One simple way to accomplish this is to write each stem value twice and then record the leaves 0, 1, 2, 3, and 4 opposite the appropriate stem value where it appears for the first time, and the leaves 5, 6, 7, 8, and 9 opposite this same stem value where it appears for the second time. This modified double-stem-and-leaf plot is illustrated in Table 1.6, where the stems corresponding to leaves 0 through 4 have been coded by the symbol and the stems corresponding to leaves 5 through 9 by the symbol ·.

In any given problem, we must decide on the appropriate stem values. This decision is made somewhat arbitrarily, although we are guided by the size of our sample. Usually, we choose between 5 and 20 stems. The smaller the number of data available, the smaller is our choice for the number of stems. For example, if

the data consist of numbers from 1 to 21 representing the number of people in a cafeteria line on 40 randomly selected workdays and we choose a double-stem-and leaf plot, the stems will be 0 , 0·, 1 , 1·, and 2 so that the smallest observation 1 has stem 0 and leaf 1, the number 18 has stem 1· and leaf 8, and the largest observation 21 has stem 2 and leaf 1. On the other hand, if the data consist of numbers from $18,800 to $19,600 representing the best possible deals on 100 new automobiles from a certain dealership and we choose a single-stem-and-leaf plot, the stems will be 188, 189, 190, ... , 196 and the leaves will now each contain two digits. A car that sold for $19,385 would have a stem value of 193 and the two-digit leaf 85. Multiple-digit leaves belonging to the same stem are usually separated by commas in the stem-and-leaf plot. Decimal points in the data are generally ignored when all the digits to the right of the decimal represent the leaf. Such was the case in Tables 1.5 and 1.6. However, if the data consist of numbers ranging from 21.8 to 74.9, we might choose the digits 2, 3, 4, 5, 6, and 7 as our stems so that a number such as 48.3 would have a stem value of 4 and a leaf of 8.3.

Table 1.6: Double-Stem-and-Leaf Plot of Battery Life

Stem	Leaf	Frequency
1·	5669	577
2 2·	00111122232	
3 3·	33444	1
4 4·	556777889	4
69	9	15 10 5
2	11234	3

The stem-and-leaf plot represents an effective way to summarize data. Another way is through the use of the frequency distribution, where the data, grouped into different classes or intervals, can be constructed by counting the leaves be longing to each stem and noting that each stem defines a class interval. In Table 1.5, the stem 1 with 2 leaves defines the interval 1.0–1.9 containing 2 observations; the stem 2 with 5 leaves defines the interval 2.0–2.9 containing 5 observations; the stem 3 with 25 leaves defines the interval 3.0–3.9 with 25 observations; and the stem 4 with 8 leaves defines the interval 4.0–4.9 containing 8 observations. For the double-stem-and-leaf plot of Table 1.6, the stems define the seven class intervals 1.5–1.9, 2.0–2.4, 2.5–2.9, 3.0–3.4, 3.5–3.9, 4.0–4.4, and 4.5–4.9 with frequencies 2, 1, 4, 15, 10, 5, and 3, respectively.

Histogram

Dividing each class frequency by the total number of observations, we obtain the proportion of the set of observations in each of the classes. A table listing relative frequencies is called a relative frequency distribution. The relative frequency distribution for the data of Table 1.4, showing the midpoint of each class interval, is given in Table 1.7.

The information provided by a relative frequency distribution in tabular form is easier to grasp if presented graphically. Using the midpoint of each interval and the

Table 1.7: Relative Frequency Distribution of Battery Life

Class Interval	Class Midpoint	Frequency, *f*	Relative Frequency
1.5–1.9	1.7	2	0.050
2.0–2.4	2.2	1	0.025
2.5–2.9	2.7	4	0.100
3.0–3.4	3.2	15	0.375
3.5–3.9	3.7	10	0.250
4.0–4.4	4.2	5	0.125
4.5–4.9	4.7	3	0.075

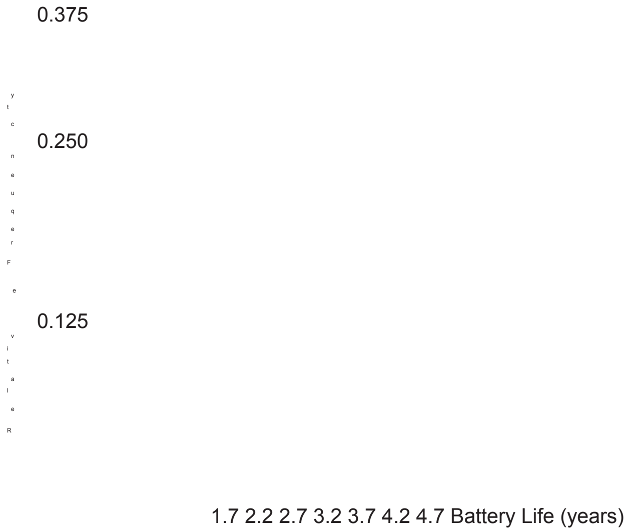

Figure 1.6: Relative frequency histogram.

corresponding relative frequency, we construct a relative frequency histogram (Figure 1.6).

Many continuous frequency distributions can be represented graphically by the characteristic bell-shaped curve of Figure 1.7. Graphical tools such as what we see in Figures 1.6 and 1.7 aid in the characterization of the nature of the population. In Chapters 5 and 6 we discuss a property of the population called its distribution. While a more rigorous definition of a distribution or probability distribution will be given later in the text, at this point one can view it as what would be seen in Figure 1.7 in the limit as the size of the sample becomes larger.

A distribution is said to be symmetric if it can be folded along a vertical axis so that the two sides coincide. A distribution that lacks symmetry with respect to a vertical axis is said to be skewed. The distribution illustrated in Figure 1.8(a) is said to be skewed to the right since it has a long right tail and a much shorter left tail. In Figure 1.8(b) we see that the distribution is symmetric, while in Figure 1.8(c) it is skewed to the left.

If we rotate a stem-and-leaf plot counterclockwise through an angle of 90°, we observe that the resulting columns of leaves form a picture that is similar to a histogram. Consequently, if our primary purpose in looking at the data is to determine the general shape or form of the distribution, it will seldom be necessary

$f(x)$

0

123456 Battery Life (years)

Figure 1.7: Estimating frequency distribution.

(a) (b) (c)

Figure 1.8: Skewness of data.

to construct a relative frequency histogram.

Box-and-Whisker Plot or Box Plot

Another display that is helpful for reflecting properties of a sample is the box and-whisker plot. This plot encloses the *interquartile range* of the data in a box that has the median displayed within. The interquartile range has as its extremes the 75th percentile (upper quartile) and the 25th percentile (lower quartile). In addition to the box, "whiskers" extend, showing extreme observations in the sample. For reasonably large samples, the display shows center of location, variability, and the degree of asymmetry.

In addition, a variation called a box plot can provide the viewer with information regarding which observations may be outliers. Outliers are observations that are considered to be unusually far from the bulk of the data. There are many statistical tests that are designed to detect outliers. Technically, one may view an outlier as being an observation that represents a "rare event" (there is a small probability of obtaining a value that far from the bulk of the data). The concept of outliers resurfaces in Chapter 12 in the context of regression analysis.

The visual information in the box-and-whisker plot or box plot is not intended to be a formal test for outliers. Rather, it is viewed as a diagnostic tool. While the determination of which observations are outliers varies with the type of software that is used, one common procedure is to use a multiple of the interquartile range. For example, if the distance from the box exceeds 1.5 times the interquartile range (in either direction), the observation may be labeled an outlier.

Example 1.5: Nicotine content was measured in a random sample of 40 cigarettes. The data are displayed in Table 1.8.

Table 1.8: Nicotine Data for Example 1.5

1.09	1.92	2.31	1.79	2.28	1.74	1.47	1.97
0.85	1.24	1.58	2.03	1.70	2.17	2.55	2.11
1.86	1.90	1.68	1.51	1.64	0.72	1.69	1.85
1.82	1.79	2.46	1.88	2.08	1.67	1.37	1.93
1.40	1.64	2.09	1.75	1.63	2.37	1.75	1.69

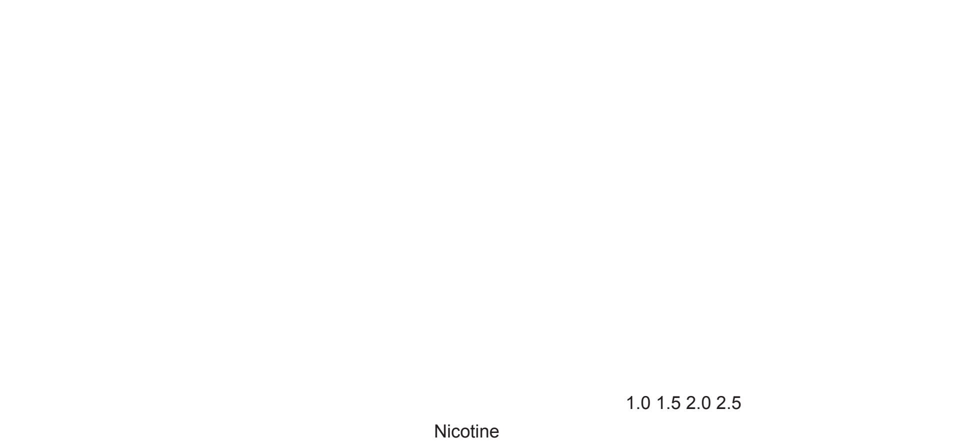

Figure 1.9: Box-and-whisker plot for Example 1.5.

Figure 1.9 shows the box-and-whisker plot of the data, depicting the observations 0.72 and 0.85 as mild outliers in the lower tail, whereas the observation 2.55 is a mild outlier in the upper tail. In this example, the interquartile range is 0.365, and 1.5 times the interquartile range is 0.5475. Figure 1.10, on the other hand, provides a stem-and-leaf plot.

Example 1.6: Consider the data in Table 1.9, consisting of 30 samples measuring the thickness of paint can “ears” (see the work by Hogg and Ledolter, 1992, in the Bibliography). Figure 1.11 depicts a box-and-whisker plot for this asymmetric set of data. Notice that the left block is considerably larger than the block on the right. The median is 35. The lower quartile is 31, while the upper quartile is 36. Notice also that the extreme observation on the right is farther away from the box than the extreme observation on the left. There are no outliers in this data set.

```
The decimal point is 1 digit(s) to the left of the |
 7|2
 8|5
 9 |
10 | 9
11 |
12 | 4
13 | 7
14 | 07
15 | 18
16 | 3447899
17 | 045599
18 | 2568
19 | 0237
20 | 389
```

21 | 17
22 | 8
23 | 17
24 | 6
25 | 5

Figure 1.10: Stem-and-leaf plot for the nicotine data.

Table 1.9: Data for Example 1.6

Sample	Measurements					Sample	Measurements				
1	29	36	39	34	34	16	35	30	35	29	37
2	29	29	28	32	31	17	40	31	38	35	31
3	34	34	39	38	37	18	35	36	30	33	32
4	35	37	33	38	41	19	35	34	35	30	36
5	30	29	31	38	29	20	35	35	31	38	36
6	34	31	37	39	36	21	32	36	36	32	36
7	30	35	33	40	36	22	36	37	32	34	34
8	28	28	31	34	30	23	29	34	33	37	35
9	32	36	38	38	35	24	36	36	35	37	37
10	35	30	37	35	31	25	36	30	35	33	31
11	35	30	35	38	35	26	35	30	29	38	35
12	38	34	35	35	31	27	35	36	30	34	36
13	34	35	33	30	34	28	35	30	36	29	35
14	40	35	34	33	35	29	38	36	35	31	31
15	34	35	38	35	30	30	30	34	40	28	30

There are additional ways that box-and-whisker plots and other graphical dis plays can aid the analyst. Multiple samples can be compared graphically. Plots of data can suggest relationships between variables. Graphs can aid in the detection of anomalies or outlying observations in samples.

There are other types of graphical tools and plots that are used. These are discussed in Chapter 8 after we introduce additional theoretical details.

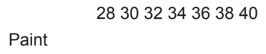

Figure 1.11: Box-and-whisker plot for thickness of paint can “ears.”

Other Distinguishing Features of a Sample

There are features of the distribution or sample other than measures of center of location and variability that further define its nature. For example, while the median divides the data (or distribution) into two parts, there are other measures that divide parts or pieces of the distribution that can be very useful. Separation is made into four parts by *quartiles*, with the third quartile separating the upper quarter of the data from the rest, the second quartile being the median, and the first quartile separating the lower quarter of the data from the rest. The distribution can be even more finely divided by computing percentiles of the distribution. These quantities give the analyst a sense of the so-called *tails* of the distribution (i.e., values that are relatively extreme, either small or large). For example, the 95th percentile separates the highest 5% from the bottom 95%. Similar definitions prevail for extremes on the lower side or *lower tail* of the distribution. The 1st percentile separates the bottom 1% from the rest of the distribution. The concept of percentiles will play a major role in much that will be covered in future chapters.

1.7 General Types of Statistical Studies: Designed Experiment, Observational Study, and Retrospective Study

In the foregoing sections we have emphasized the notion of sampling from a population and the use of statistical methods to learn or perhaps affirm important information about the population. The information sought and learned through the use of these statistical methods can often be influential in decision making and problem solving in many important scientific and engineering areas. As an illustra tion, Example 1.3 describes a simple experiment in which the results may provide an aid in determining the kinds of conditions under which it is not advisable to use a particular aluminum alloy that may have a dangerous vulnerability to corrosion. The results may be of use not only to those who produce the alloy, but also to the customer who may consider using it. This illustration, as well as many more that appear in Chapters 13 through 15, highlights the concept of designing or control ling experimental conditions (combinations of coating conditions and humidity) of interest to learn about some characteristic or measurement (level of corrosion) that results from these conditions. Statistical methods that make use of measures of central tendency in the corrosion measure, as well as measures of variability, are employed. As the reader will observe later in the text, these methods often lead to a statistical model like that discussed in Section 1.6. In this case, the model may be used to estimate (or predict) the corrosion measure as a function of humidity and the type of coating employed. Again, in developing this kind of model, descriptive statistics that highlight central tendency and variability become very useful.

The information supplied in Example 1.3 illustrates nicely the types of engineering questions asked and answered by the use of statistical methods that are employed through a designed experiment and presented in this text. They are

(i) What is the nature of the impact of relative humidity on the corrosion of the aluminum alloy within the range of relative humidity in this experiment?

(ii) Does the chemical corrosion coating reduce corrosion levels and can the effect be quantified in some fashion?

(iii) Is there interaction between coating type and relative humidity that impacts their influence on corrosion of the alloy? If so, what is its interpretation?

What Is Interaction?

The importance of questions (i) and (ii) should be clear to the reader, as they deal with issues important to both producers and users of the alloy. But what about question (iii)? The concept of *interaction* will be discussed at length in Chapters 14 and 15. Consider the plot in Figure 1.3. This is an illustration of the detection of interaction between two factors in a simple designed experiment. Note that the lines connecting the sample means are not parallel. Parallelism would have indicated that the effect (seen as a result of the slope of the lines) of relative humidity is the same, namely a negative effect, for both an uncoated condition and the chemical corrosion coating. Recall that the negative slope implies that corrosion becomes more pronounced as humidity rises. Lack of parallelism implies an interaction between coating type and relative humidity. The nearly “flat” line for the corrosion coating as opposed to a steeper slope for the uncoated condition suggests that *not only is the chemical corrosion coating beneficial (note the displacement between the lines), but the presence of the coating renders the effect of humidity negligible*. Clearly all these questions are very important to the effect of the two individual factors and to the interpretation of the interaction, if it is present.

Statistical models are extremely useful in answering questions such as those listed in (i), (ii), and (iii), where the data come from a designed experiment. But one does not always have the luxury or resources to employ a designed experiment. For example, there are many instances in which the conditions of interest to the scientist or engineer cannot be implemented simply because the *important factors cannot be controlled*. In Example 1.3, the relative humidity and coating type (or lack of coating) are quite easy to control. This of course is the defining feature of a designed experiment. In many fields, factors that need to be studied cannot be controlled for any one of various reasons. Tight control as in Example 1.3 allows the analyst to be confident that any differences found (for example, in corrosion levels) are due to the factors under control. As a second illustration, consider Exercise 1.6 on page 33. Suppose in this case 24 specimens of silicone rubber are selected and 12 assigned to each of the curing temperature levels. The temperatures are controlled carefully, and thus this is an example of a designed experiment with a single factor being curing temperature. Differences found in the mean tensile strength would be assumed to be attributed to the different curing temperatures.

What If Factors Are Not Controlled?

Suppose there are no factors controlled and *no random assignment* of fixed treatments to experimental units and yet there is a need to glean information from a data set. As an illustration, consider a study in which interest centers around the relationship between blood cholesterol levels and the amount of sodium measured in the blood. A group of individuals were monitored over time for both blood cholesterol and sodium. Certainly some useful information can be gathered from such a data set. However, it should be clear that there certainly can be no strict control of blood sodium levels. Ideally, the subjects should be divided randomly into two groups, with one group assigned a specific high level of blood sodium and the other a specific low level of blood sodium. Obviously this cannot be done. Clearly changes in cholesterol can be experienced because of changes in one of a number of other factors that were not controlled. This kind of study, without factor control, is called an observational study. Much of the time it involves a situation in which subjects are observed across time.

Biological and biomedical studies are often by necessity observational studies. However, observational studies are not confined to those areas. For example, con sider a study that is designed to determine the influence of ambient temperature on the electric power consumed by a chemical plant. Clearly, levels of ambient temper ature cannot be controlled, and thus the data structure can only be a monitoring of the data from the plant over time.

It should be apparent that the striking difference between a well-designed ex periment and observational studies is the difficulty in determination of true cause and effect with the latter. Also, differences found in the fundamental response (e.g., corrosion levels, blood cholesterol, plant electric power consumption) may be due to other underlying factors that were not controlled. Ideally, in a designed experiment the *nuisance factors* would be equalized via the randomization process. Certainly changes in blood cholesterol could be due to fat intake, exercise activity, and so on. Electric power consumption could be affected by the amount of product produced or even the purity of the product produced.

Another often ignored disadvantage of an observational study when compared to carefully designed experiments is that, unlike the latter, observational studies are at the mercy of nature, environmental or other uncontrolled circumstances that impact the ranges of factors of interest. For example, in the biomedical study regarding the influence of blood sodium levels on blood cholesterol, it is possible that there is indeed a strong influence but the particular data set used did not involve enough observed variation in sodium levels because of the nature of the subjects chosen. Of course, in a designed experiment, the analyst chooses and controls ranges of factors.

A third type of statistical study which can be very useful but has clear dis advantages when compared to a designed experiment is a retrospective study. This type of study uses strictly historical data, data taken over a specific period of time. One obvious advantage of retrospective data is that there is reduced cost in collecting the data. However, as one might expect, there are clear disadvantages.

(i) Validity and reliability of historical data are often in doubt.

(ii) If time is an important aspect of the structure of the data, there may be data missing.

(iii) There may be errors in collection of the data that are not known.

(iv) Again, as in the case of observational data, there is no control on the ranges of the measured variables (the factors in a study). Indeed, the ranges found in historical data may not be relevant for current studies.

In Section 1.6, some attention was given to modeling of relationships among vari ables. We introduced the notion of regression analysis, which is covered in Chapters 11 and 12 and is illustrated as a form of data analysis for designed experiments discussed in Chapters 14 and 15. In Section 1.6, a model relating population mean tensile strength of cloth to percentages of cotton was used for illustration, where 20 specimens of cloth represented the experimental units. In that case, the data came from a simple designed experiment where the individual cotton percentages were selected by the scientist.

Often both observational data and retrospective data are used for the purpose of observing relationships among variables through model-building procedures dis cussed in Chapters 11 and 12. While the advantages of designed experiments certainly apply when the goal is statistical model building, there are many areas in which designing of experiments is not possible. Thus, *observational or historical data must be used*. We refer here to a historical data set that is found in Exercise 12.5 on page 472. The goal is to build a model that will result in an equation or relationship that relates monthly electric power consumed to average ambient temperature x_1, the number of days in the month x_2, the average product purity x_3, and the tons of product produced x_4. The data are the past year's historical data.

Exercises

1.13 A manufacturer of electronic components is interested in determining the lifetime of a certain type of battery. A sample, in hours of life, is as follows:

123, 116, 122, 110, 192, 126, 125, 111, 118, 117.

(a) Find the sample mean and median.

(b) What feature in this data set is responsible for the substantial difference between the two?

1.14 A tire manufacturer wants to determine the inner diameter of a certain grade of tire. Ideally, the diameter would be 570 mm. The data are as follows:

572, 572, 573, 568, 569, 575, 565, 570.

(a) Find the sample mean and median.

(b) Find the sample variance, standard deviation, and range.

(c) Using the calculated statistics in parts (a) and (b), can you comment on the quality of the tires?

1.15 Six independent coin tosses result in *HHHHHH*. It turns out that if the coin is fair, the probability of this outcome is $(1/2)^6 = 0.015625$. Does this produce strong evidence that the coin is not fair? Comment and use the concept of *P*-value discussed in Section 1.1.

(c) Compute the sample mean, sample range, and sam

1.16 Show that the *n* pieces of information in

^{n} $_{i=1}$

$(x_i - \bar{x})=0.$

$_{i=1}$

length of life, in seconds, of 50 fruit flies subject to a new spray in a controlled laboratory experiment:

$(x_i - \bar{x})^2$ are not independent; that is, show that

ple standard deviation.

n

1.20 The following data represent the

1.17 A study of the effects of smoking on sleep patterns is conducted. The measure observed is the time, in minutes, that it takes to fall asleep. These data are obtained:

Smokers: 69.3 56.0 22.1 47.6 53.2 48.1
52.7 34.4
60.2 43.8 23.2 13.8
Nonsmokers: 28.6 25.1 26.4 34.9 29.8
28.4 38.5 30.2
30.6 31.8 41.6 21.1
36.0 37.9 13.9

(a) Find the sample mean for each group. (b) Find the sample standard deviation for each group. (c) Make a dot plot of the data sets A and B on the same line.

(d) Comment on what kind of impact smoking appears to have on the time required to fall asleep.

1.18 The following scores represent the final examination grades for an elementary statistics course: 23 60 79 32 57 74 52 70 82 36 80 77 81 95 41 65 92 85 55 76 52 10 64 75 78 25 80 98 81 67 41 71 83 54 64 72 88 62 74 43 60 78 89 76 84 48 84 90 15 79 34 67 17 82 69 74 63 80 85 61

(a) Construct a stem-and-leaf plot for the examination grades in which the stems are 1, 2, 3,..., 9. (b) Construct a relative frequency histogram, draw an estimate of the graph of the distribution, and discuss the skewness of the distribution.

(c) Compute the sample mean, sample median, and sample standard deviation.

1.19 The following data represent the length of life in years, measured to the nearest tenth, of 30 similar fuel pumps:

2.0 3.0 0.3 3.3 1.3 0.4
0.2 6.0 5.5 6.5 0.2 2.3
1.5 4.0 5.9 1.8 4.7 0.7
4.5 0.3 1.5 0.5 2.5 5.0
1.0 6.0 5.6 6.0 1.2 0.2

(a) Construct a stem-and-leaf plot for the life in years of the fuel pumps, using the digit to the left of the decimal point as the stem for each observation. (b) Set up a relative frequency distribution.

17 20 10 9 23 13 12 19 18 24 12 14 6 9 13 6 7 10
13 7 16 18 8 13 3 32 9 7 10 11 13 7 18 7 10 4 27
19 16 8 7 10 5 14 15 10 9 6 7 15

(a) Construct a double-stem-and-leaf plot for the life span of the fruit flies using the stems 0 , 0·, 1 , 1·, 2 , 2·, and 3 such that stems coded by the symbols and · are associated, respectively, with leaves 0 through 4 and 5 through 9.

(b) Set up a relative frequency distribution. (c) Construct a relative frequency histogram. (d) Find the median.

1.21 The lengths of power failures, in minutes, are recorded in the following table.

22 18 135 83 55 28 70 66 74 40 98 87 50 96
118 15 90 78 121 120 13 89 103 24 132 115
21 158 74 78 69 22 21 28 83 98 102 124 112
120 121 43 37 93 95

(a) Find the sample mean and sample median of the power-failure times.

(b) Find the sample standard deviation of the power failure times.

1.22 The following data are the measures of the diameters of 36 rivet heads in 1/100 of an inch.

6.72 6.77 6.82 6.70 6.78 6.70 6.62 6.75 6.66
6.66 6.64 6.76 6.73 6.80 6.72 6.76 6.76 6.68
6.66 6.62 6.72 6.76 6.70 6.78 6.76 6.67 6.70
6.72 6.74 6.81 6.79 6.78 6.66 6.76 6.76 6.72

(a) Compute the sample mean and sample standard deviation.

(b) Construct a relative frequency histogram of the data.

(c) Comment on whether or not there is any clear indication that the sample came from a population that has a bell-shaped distribution.

1.23 The hydrocarbon emissions at idling speed in parts per million (ppm) for automobiles of 1980 and 1990 model years are given for 20 randomly selected cars.

1980 models:
141 359 247 940 882 494 306 210 105 880
200 223 188 940 241 190 300 435 241 380
1990 models:
140 160 20 20 223 60 20 95 360 70 220 400
217 58 235 380 200 175 85 65

(a) Construct a dot plot as in Figure 1.1. (b) Compute the sample means for the two years and superimpose the two means on the plots. (c) Comment on what the dot plot indicates regarding whether or not the population emissions changed from 1980 to 1990. Use the concept of variability in your comments.

1.24 The following are historical data on staff salaries (dollars per pupil) for 30 schools sampled in the eastern part of the United States in the early 1970s.

3.79 2.99 2.77 2.91 3.10 1.84 2.52 3.22 2.45
2.14 2.67 2.52 2.71 2.75 3.57 3.85 3.36 2.05
2.89 2.83 3.13 2.44 2.10 3.71 3.14 3.54 2.37
2.68 3.51 3.37

(a) Compute the sample mean and sample standard deviation.

(b) Construct a relative frequency histogram of the data.

(c) Construct a stem-and-leaf display of the data.

1.25 The following data set is related to that in Exercise 1.24. It gives the percentages of the families that are in the upper income level, for the same individual schools in the same order as in Exercise 1.24.

72.2 31.9 26.5 29.1 27.3 8.6 22.3 26.5
20.4 12.8 25.1 19.2 24.1 58.2 68.1 89.2
55.1 9.4 14.5 13.9 20.7 17.9 8.5 55.4 38.1
54.2 21.5 26.2 59.1 43.3

(a) Calculate the sample mean.

(b) Calculate the sample median.

(c) Construct a relative frequency histogram of the data.

(d) Compute the 10% trimmed mean. Compare with the results in (a) and (b) and comment.

1.26 Suppose it is of interest to use the data sets in Exercises 1.24 and 1.25 to derive a model that would predict staff salaries as a function of percentage of families in a high income level for current school systems. Comment on any disadvantage in carrying out this type of analysis.

1.27 A study is done to determine the influence of the wear, *y*, of a bearing as a function of the load, *x*, on the bearing. A designed experiment is used for this study. Three levels of load were used, 700 lb, 1000 lb, and 1300 lb. Four specimens were used at each level, and the sample means were, respectively, 210, 325, and 375.

(a) Plot average wear against load.

(b) From the plot in (a), does it appear as if a relationship exists between wear and load?

(c) Suppose we look at the individual wear values for each of the four specimens at each load level (see the data that follow). Plot the wear results for all specimens against the three load values.

(d) From your plot in (c), does it appear as if a clear relationship exists? If your answer is different from that in (b), explain why.

	x		
	700	1000	1300
~~y_1~~	~~145~~	~~250~~	~~150~~
y_2	105	195	180
y_3	260	375	420
y_4	330	480	750

$\bar{y}_1 = 210$ $\bar{y}_2 = 325$ $\bar{y}_3 = 375$

1.28 Many manufacturing companies in the United States and abroad use molded parts as components of a process. Shrinkage is often a major problem. Thus, a molded die for a part is built larger than nominal size to allow for part shrinkage. In an injection molding study it is known that the shrinkage is influenced by many factors, among which are the injection velocity in ft/sec and mold temperature in °C. The following two data sets show the results of a designed experiment in which injection velocity was held at two levels (low and high) and mold temperature was held constant at a low level. The shrinkage is measured in cm × 10^4 .

Shrinkage values at low injection velocity:

72.68 72.62 72.58 72.48 73.07
72.55 72.42 72.84 72.58 72.92

Shrinkage values at high injection velocity:

71.62 71.68 71.74 71.48 71.55
71.52 71.71 71.56 71.70 71.50

(a) Construct a dot plot of both data sets on the same graph. Indicate on the plot both shrinkage means, that for low injection velocity and high injection velocity.

(b) Based on the graphical results in (a), using the location of the two means and your sense of variability, what do you conclude regarding the effect of injection velocity on shrinkage at low mold temperature?

1.29 Use the data in Exercise 1.24 to construct a box plot.

1.30 Below are the lifetimes, in hours, of fifty 40-watt, 110-volt internally frosted incandescent lamps, taken from forced life tests:

919 1196 785 1126 936 918 1156 920
948 1067 1092 1162 1170 929 950 905
972 1035 1045 855 1195 1195 1340 1122
938 970 1237 956 1102 1157 978 832
1009 1157 1151 1009 765 958 902 1022
1333 811
1217 1085 896 958 1311 1037 702 923

Construct a box plot for these data.

1.31 Consider the situation of Exercise 1.28. But now use the following data set, in which shrinkage is measured once again at low injection velocity and high injection velocity. However, this time the mold temperature is raised to a high level and held constant.

Shrinkage values at low injection velocity:

76.20 76.09 75.98 76.15 76.17
75.94 76.12 76.18 76.25 75.82

Shrinkage values at high injection velocity:

93.25 93.19 92.87 93.29 93.37
92.98 93.47 93.75 93.89 91.62

(a) As in Exercise 1.28, construct a dot plot with both data sets on the same graph and identify both means (i.e., mean shrinkage for low injection velocity and for high injection velocity).

(b) As in Exercise 1.28, comment on the influence of injection velocity on shrinkage for high mold temperature. Take into account the position of the two means and the variability around each mean.

(c) Compare your conclusion in (b) with that in (b) of Exercise 1.28 in which mold temperature was held at a low level. Would you say that there is an interaction between injection velocity and mold temperature? Explain.

1.32 Use the results of Exercises 1.28 and 1.31 to create a plot that illustrates the interaction evident from the data. Use the plot in Figure 1.3 in Example 1.3 as a guide. Could the type of information found in Exercises 1.28 and 1.31 have been found in an observational study in which there was no control on injection veloc ity and mold temperature by the analyst? Explain why or why not.

1.33 Group Project: Collect the data for the height of everyone in the class. Use the sample means and variances and the types of plots presented in this chapter to summarize any features that draw a distinction between the distributions of heights for males and females. Do the same for the weight of everyone in the class.

This page intentionally left blank

Chapter 2

Probability

2.1 Sample Space

In the study of statistics, we are concerned basically with the presentation and interpretation of chance outcomes that occur in a planned study or scientific investigation. For example, we may record the number of accidents that occur monthly at the intersection of Driftwood Lane and Royal Oak Drive, hoping to justify the installation of a traffic light; we might classify items coming off an assembly line as “defective” or “nondefective”; or we may be interested in the volume of gas released in a chemical reaction when the concentration of an acid is varied. Hence, the statistician is often dealing with either numerical data, representing counts or measurements, or categorical data, which can be classified according to some criterion.

We shall refer to any recording of information, whether it be numerical or categorical, as an observation. Thus, the numbers 2, 0, 1, and 2, representing the number of accidents that occurred for each month from January through April during the past year at the intersection of Driftwood Lane and Royal Oak Drive, constitute a set of observations. Similarly, the categorical data *N, D, N, N,* and *D*, representing the items found to be defective or nondefective when five items are inspected, are recorded as observations.

Statisticians use the word experiment to describe any process that generates a set of data. A simple example of a statistical experiment is the tossing of a coin. In this experiment, there are only two possible outcomes, heads or tails. Another experiment might be the launching of a missile and observing of its velocity at specified times. The opinions of voters concerning a new sales tax can also be considered as observations of an experiment. We are particularly interested in the observations obtained by repeating the experiment several times. In most cases, the outcomes will depend on chance and, therefore, cannot be predicted with certainty. If a chemist runs an analysis several times under the same conditions, he or she will obtain different measurements, indicating an element of chance in the experimental procedure. Even when a coin is tossed repeatedly, we cannot be certain that a given toss will result in a head. However, we know the entire set of possibilities for each toss.

Given the discussion in Section 1.7, we should deal with the breadth of the term *experiment*. Three types of statistical studies were reviewed, and several examples were given of each. In each of the three cases, *designed experiments*, *observational studies*, and *retrospective studies*, the end result was a set of *data* that of course is

subject to uncertainty. Though only one of these has the word *experiment* in its description, the process of generating the data or the process of observing the data is part of an experiment. The corrosion study discussed in Section 1.2 certainly involves an experiment, with measures of corrosion representing the data. The ex ample given in Section 1.7 in which blood cholesterol and sodium were observed on a group of individuals represented an observational study (as opposed to a *designed* experiment), and yet the process generated data and the outcome is subject to un certainty. Thus, it is an experiment. A third example in Section 1.7 represented a retrospective study in which historical data on monthly electric power consump tion and average monthly ambient temperature were observed. Even though the data may have been in the files for decades, the process is still referred to as an experiment.

Definition 2.1:

The set of all possible outcome
sample space and is represent

Each outcome in a sample space is called an element or a member of the sample space, or simply a sample point. If the sample space has a finite number of elements, we may *list* the members separated by commas and enclosed in braces. Thus, the sample space *S*, of possible outcomes when a coin is flipped, may be written where H and T correspond to heads and tails, respectively.

$S = \{H, T\}$,

Example 2.1: Consider the experiment of tossing a die. If we are interested in the number that shows on the top face, the sample space is

$$S_1 = \{1, 2, 3, 4, 5, 6\}.$$

If we are interested only in whether the number is even or odd, the sample space is simply

$$S_2 = \{\text{even}, \text{odd}\}.$$

Example 2.1 illustrates the fact that more than one sample space can be used to describe the outcomes of an experiment. In this case, S_1 provides more information than S_2 . If we know which element in S_1 occurs, we can tell which outcome in S_2 occurs; however, a knowledge of what happens in S_2 is of little help in determining which element in S_1 occurs. In general, it is desirable to use the sample space that gives the most information concerning the outcomes of the experiment. In some experiments, it is helpful to list the elements of the sample space systematically by means of a tree diagram.

Example 2.2: An experiment consists of flipping a coin and then flipping it a second time if a head occurs. If a tail occurs on the first flip, then a die is tossed once. To list the elements of the sample space providing the most information, we construct the tree diagram of Figure 2.1. The various paths along the branches of the tree give the distinct sample points. Starting with the top left branch and moving to the right along the first path, we get the sample point HH, indicating the possibility that heads occurs on two successive flips of the coin. Likewise, the sample point $T3$ indicates the possibility that the coin will show a tail followed by a 3 on the toss of the die. By proceeding along all paths, we see that the sample space is

$$S = \{HH, HT, T1, T2, T3, T4, T5, T6\}.$$

Outcome

Sampl
e Point

Second
Outcom
e

First

H HH T HT

H

1

T 1 *T* 2 *T* 3

2 3 4 5 6

T 4 *T* 5 *T* 6

T

Figure 2.1: Tree diagram for Example 2.2.

Many of the concepts in this chapter are best illustrated with examples involving the use of dice and cards. These are particularly important applications to use early in the learning process, to facilitate the flow of these new concepts into scientific and engineering examples such as the following.

Example 2.3: Suppose that three items are selected at random from a manufacturing process. Each item is inspected and classified defective, *D*, or nondefective, *N*. To list the elements of the sample space providing the most information, we construct the tree diagram of Figure 2.2. Now, the various paths along the branches of the tree give the distinct sample points. Starting with the first path, we get the sample point *DDD*, indicating the possibility that all three items inspected are defective. As we proceed along the other paths, we see that the sample space is

$$S = \{DDD, DDN, DND, DNN, NDD, NDN, NND, NNN\}.$$

Sample spaces with a large or infinite number of sample points are best described by a statement or rule method. For example, if the possible outcomes of an experiment are the set of cities in the world with a population over 1 million, our sample space is written

$$S = \{x \mid x \text{ is a city with a population over 1 million}\},$$

which reads "*S* is the set of all *x* such that *x* is a city with a population over 1 million." The vertical bar is read "such that." Similarly, if *S* is the set of all points (*x*, *y*) on the boundary or the interior of a circle of radius 2 with center at the origin, we write the rule

$$S = \{(x, y) \mid x^2 + y^2 \leq 4\}.$$

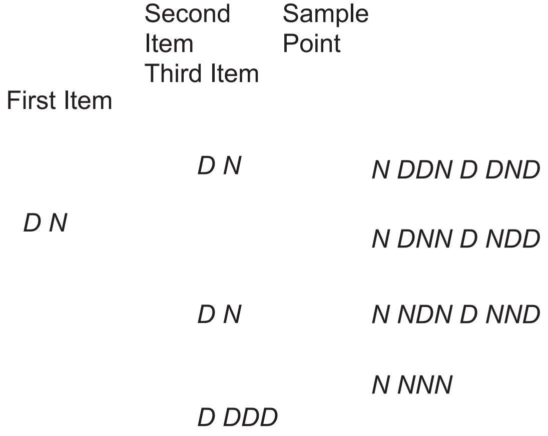

Figure 2.2: Tree diagram for Example 2.3.

Whether we describe the sample space by the rule method or by listing the elements will depend on the specific problem at hand. The rule method has practical advantages, particularly for many experiments where listing becomes a tedious chore.

Consider the situation of Example 2.3 in which items from a manufacturing process are either *D*, defective, or *N*, nondefective. There are many important statistical procedures called sampling plans that determine whether or not a "lot" of items is considered satisfactory. One such plan involves sampling until *k* defectives are observed. Suppose the experiment is to sample items randomly until one defective item is observed. The sample space for this case is

$$S = \{D, ND, NND, NNND, \dots \}.$$

2.2 Events

For any given experiment, we may be interested in the occurrence of certain events rather than in the occurrence of a specific element in the sample space. For in stance, we may be interested in the event A that the outcome when a die is tossed is divisible by 3. This will occur if the outcome is an element of the subset $A = \{3, 6\}$ of the sample space S_1 in Example 2.1. As a further illustration, we may be interested in the event B that the number of defectives is greater than 1 in Example 2.3. This will occur if the outcome is an element of the subset

$$B = \{DDN, DND, NDD, DDD\}$$

of the sample space S.

To each event we assign a collection of sample points, which constitute a subset of the sample space. That subset represents all of the elements for which the event is true.

Definition 2.2:

An event is a subset of a samp

Example 2.4: Given the sample space $S = \{t \mid t \geq 0\}$, where t is the life in years of a certain electronic component, then the event A that the component fails before the end of the fifth year is the subset $A = \{t \mid 0 \leq t < 5\}$.

It is conceivable that an event may be a subset that includes the entire sample space S or a subset of S called the null set and denoted by the symbol φ, which contains no elements at all. For instance, if we let A be the event of detecting a microscopic organism by the naked eye in a biological experiment, then $A = \varphi$. Also, if

$$B = \{x \mid x \text{ is an even factor of } 7\},$$

then B must be the null set, since the only possible factors of 7 are the odd numbers 1 and 7.

Consider an experiment where the smoking habits of the employees of a man ufacturing firm are recorded. A possible sample space might classify an individual as a nonsmoker, a light smoker, a moderate smoker, or a heavy smoker. Let the subset of smokers be some event. Then all the nonsmokers correspond to a different event, also a subset of S, which is called the complement of the set of smokers.

Definition 2.3:

The complement of an event A
of S that are not in A. We denc

Example 2.5: Let R be the event that a red card is selected from an ordinary deck of 52 playing cards, and let S be the entire deck. Then R is the event that the card selected from the deck is not a red card but a black card.

Example 2.6: Consider the sample space

$$S = \{\text{book, cell phone, mp3, paper, stationery, laptop}\}.$$

Let A = {book, stationery, laptop, paper}. Then the complement of A is A = {cell phone, mp3}.

We now consider certain operations with events that will result in the formation of new events. These new events will be subsets of the same sample space as the given events. Suppose that A and B are two events associated with an experiment. In other words, A and B are subsets of the same sample space S. For example, in the tossing of a die we might let A be the event that an even number occurs and B the event that a number greater than 3 shows. Then the subsets $A = \{2, 4, 6\}$ and $B = \{4, 5, 6\}$ are subsets of the same sample space

$$S = \{1, 2, 3, 4, 5, 6\}.$$

Note that *both* A and B will occur on a given toss if the outcome is an element of the subset {4, 6}, which is just the intersection of A and B.

Definition 2.4:

The intersection of two events
event containing all elements t

Example 2.7: Let E be the event that a person selected at random in a classroom is majoring in engineering, and let F be the event that the person is female. Then $E \cap F$ is the event of all female engineering students in the classroom.

Example 2.8: Let $V = \{a, e, i, o, u\}$ and $C = \{l, r, s, t\}$; then it follows that $V \cap C = \varphi$. That is, V and C have no elements in common and, therefore, cannot both simultaneously occur.

For certain statistical experiments it is by no means unusual to define two events, A and B, that cannot both occur simultaneously. The events A and B are then said to be mutually exclusive. Stated more formally, we have the following definition:

Definition 2.5:

Two events A and B are mutua
A and B have no elements in c

Example 2.9: A cable television company offers programs on eight different channels, three of which are affiliated with ABC, two with NBC, and one with CBS. The other two are an educational channel and the ESPN sports channel. Suppose that a person subscribing to this service turns on a television without first selecting the channel. Let A be the event that the program belongs to the NBC network and B the event that it belongs to the CBS network. Since a television program cannot belong to more than one network, the events A and B have no programs in common. Therefore, the intersection $A \cap B$ contains no programs, and consequently the events A and B are mutually exclusive.

Often one is interested in the occurrence of at least one of two events associated with an experiment. Thus, in the die-tossing experiment, if

$$A = \{2, 4, 6\} \text{ and } B = \{4, 5, 6\},$$

we might be interested in either A or B occurring or both A and B occurring. Such an event, called the union of A and B, will occur if the outcome is an element of the subset {2, 4, 5, 6}.

Definition 2.6:

The union of the two events *A*
event containing all the elemer

Example 2.10: Let *A* = *{a, b, c}* and *B* = *{b, c, d, e}*; then *A* ∪ *B* = *{a, b, c, d, e}*.

Example 2.11: Let *P* be the event that an employee selected at random from an oil drilling company smokes cigarettes. Let *Q* be the event that the employee selected drinks alcoholic beverages. Then the event *P* ∪ *Q* is the set of all employees who either drink or smoke or do both.

Example 2.12: If $M = \{x \mid 3 < x < 9\}$ and $N = \{y \mid 5 < y < 12\}$, then

$$M \cup N = \{z \mid 3 < z < 12\}.$$

The relationship between events and the corresponding sample space can be illustrated graphically by means of Venn diagrams. In a Venn diagram we let the sample space be a rectangle and represent events by circles drawn inside the rectangle. Thus, in Figure 2.3, we see that

$$A \cap B = \text{regions 1 and 2},$$
$$B \cap C = \text{regions 1 and 3},$$

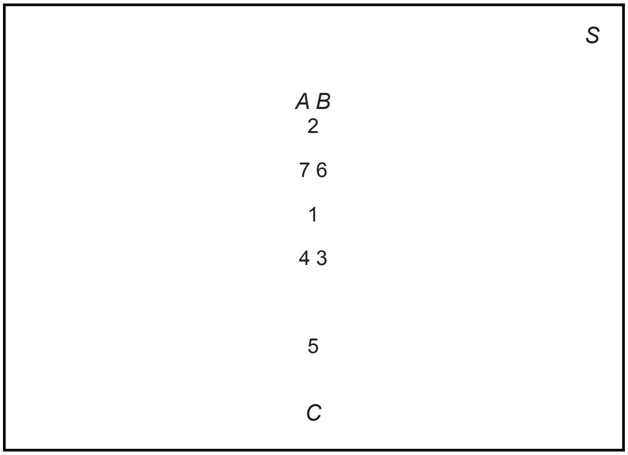

Figure 2.3: Events represented by various regions.

$$A \cup C = \text{regions 1, 2, 3, 4, 5, and 7},$$
$$B \cap A = \text{regions 4 and 7},$$
$$A \cap B \cap C = \text{region 1},$$
$$(A \cup B) \cap C = \text{regions 2, 6, and 7},$$

and so forth.

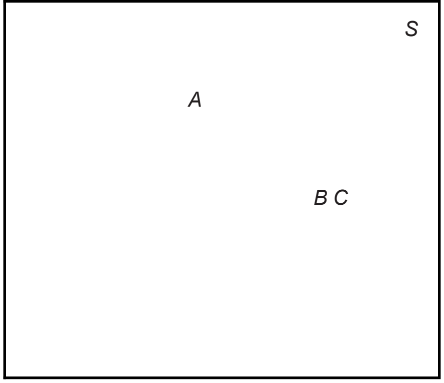

Figure 2.4: Events of the sample space S.

In Figure 2.4, we see that events A, B, and C are all subsets of the sample space S. It is also clear that event B is a subset of event A; event $B \cap C$ has no elements and hence B and C are mutually exclusive; event $A \cap C$ has at least one element; and event $A \cup B = A$. Figure 2.4 might, therefore, depict a situation where we select a card at random from an ordinary deck of 52 playing cards and observe whether the following events occur:

A: the card is red,

B: the card is the jack, queen, or king of diamonds,

C: the card is an ace.

Clearly, the event $A \cap C$ consists of only the two red aces.

Several results that follow from the foregoing definitions, which may easily be verified by means of Venn diagrams, are as follows:

1. $A \cap \varphi = \varphi$. 2. $A \cup \varphi = A$. 3. $A \cap A\ = \varphi$.

4. $A \cup A\ = S$. 5. $S\ = \varphi$.

6. $\varphi\ = S$.

7. $(A\ )\ = A$.

8. $(A \cap B)\ = A\ \cup B\ $.

9. $(A \cup B)\ = A\ \cap B\ $.

Exercises

2.1 List the elements of each of the following sample spaces:

(a) the set of integers between 1 and 50 divisible by 8;

(b) the set $S = \{x \mid x^2 + 4x - 5=0\}$;

(c) the set of outcomes when a coin is tossed until a tail or three heads appear;

(d) the set $S = \{x \mid x$ is a continent$\}$;

(e) the set $S = \{x \mid 2x - 4 \geq 0$ and $x < 1\}$.

2.2 Use the rule method to describe the sample space S consisting of all points in the first quadrant inside a circle of radius 3 with center at the origin.

2.3 Which of the following events are equal?

(a) $A = \{1, 3\}$;

(b) $B = \{x \mid x$ is a number on a die$\}$;

(c) $C = \{x \mid x^2 - 4x + 3=0\}$;

(d) $D = \{x \mid x$ is the number of heads when six coins are tossed$\}$.

2.4 An experiment involves tossing a pair of dice, one green and one red, and recording the numbers that come up. If x equals the outcome on the green die and y the outcome on the red die, describe the sample space S

(a) by listing the elements (x, y);

(b) by using the rule method.

2.5 An experiment consists of tossing a die and then flipping a coin once if the number on the die is even. If the number on the die is odd, the coin is flipped twice. Using the notation $4H$, for example, to denote the out come that the die comes up 4 and then the coin comes up heads, and 3*HT* to denote the outcome that the die comes up 3 followed by a head and then a tail on the coin, construct a tree diagram to show the 18 elements of the sample space *S*.

2.6 Two jurors are selected from 4 alternates to serve at a murder trial. Using the notation A_1A_3 , for exam ple, to denote the simple event that alternates 1 and 3 are selected, list the 6 elements of the sample space *S*.

2.7 Four students are selected at random from a chemistry class and classified as male or female. List the elements of the sample space S_1 , using the letter *M* for male and *F* for female. Define a second sample space S_2 where the elements represent the number of females selected.

2.8 For the sample space of Exercise 2.4, (a) list the elements corresponding to the event *A* that the sum is greater than 8;

(b) list the elements corresponding to the event *B* that a 2 occurs on either die;

(c) list the elements corresponding to the event *C* that a number greater than 4 comes up on the green die;

(d) list the elements corresponding to the event $A \cap C$;

(e) list the elements corresponding to the event $A \cap B$;

(f) list the elements corresponding to the event $B \cap C$;

(g) construct a Venn diagram to illustrate the intersec tions and unions of the events *A*, *B*, and *C*.

2.9 For the sample space of Exercise 2.5, (a) list the elements corresponding to the event *A* that a number less than 3 occurs on the die;

(b) list the elements corresponding to the event *B* that two tails occur;

(c) list the elements corresponding to the event *A* ;

(d) list the elements corresponding to the event $A \cap B$;

(e) list the elements corresponding to the event $A \cup B$.

2.10 An engineering firm is hired to determine if cer tain waterways in Virginia are safe for fishing. Samples are taken from three rivers.

(a) List the elements of a sample space *S*, using the letters *F* for safe to fish and *N* for not safe to fish. (b) List the elements of *S* corresponding to event *E* that at least two of the rivers are safe for fishing. (c) Define an event that has as its elements the points

{FFF, NFF, FFN, NFN}.

2.11 The resum´es of two male applicants for a college teaching position in chemistry are placed in the same file as the resum´es of two female applicants. Two po sitions become available, and the first, at the rank of assistant professor, is filled by selecting one of the four applicants at random. The second position, at the rank of instructor, is then filled by selecting at random one of the remaining three applicants. Using the notation M_2F_1 , for example, to denote the simple event that the first position is filled by the second male applicant and the second position is then filled by the first female applicant,

(a) list the elements of a sample space *S*; (b) list the elements of *S* corresponding to event *A* that the position of assistant professor is filled by a male applicant;

(c) list the elements of *S* corresponding to event *B* that exactly one of the two positions is filled by a male applicant;

(d) list the elements of *S* corresponding to event *C* that neither position is filled by a male applicant; (e) list the elements of *S* corresponding to the event $A \cap B$;

(f) list the elements of *S* corresponding to the event $A \cup C$;

(g) construct a Venn diagram to illustrate the intersec tions and unions of the events *A*, *B*, and *C*.

2.12 Exercise and diet are being studied as possi ble substitutes for medication to lower blood pressure. Three groups of subjects will be used to study the ef fect of exercise. Group 1 is sedentary, while group 2 walks and group 3 swims for 1 hour a day. Half of each of the three exercise groups will be on a salt-free diet. An additional group of subjects will not exercise or re strict their salt, but will take the standard medication. Use *Z* for sedentary, *W* for walker, *S* for swimmer, *Y* for salt, *N* for no salt, *M* for medication, and *F* for medication free.

(a) Show all of the elements of the sample space *S*.

(b) Given that *A* is the set of nonmedicated subjects and *B* is the set of walkers, list the elements of $A \cup B$.

(c) List the elements of $A \cap B$.

2.13 Construct a Venn diagram to illustrate the pos sible intersections and unions for the following events relative to the sample space consisting of all automo biles made in the United States.

F : Four door, *S* : Sun roof, *P* : Power steering.

2.14 If $S = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$, $A = \{0, 2, 4, 6, 8\}$, $B = \{1, 3, 5, 7, 9\}$, $C = \{2, 3, 4, 5\}$, and $D = \{1, 6, 7\}$, list the elements of the sets correspond ing to the following events:

(a) $A \cup D$;

(b) $A \cap B$;

(c) *B* ;

(d) $(B \cap D) \cup A$;

(e) $(S \cap D)$;

(f) $A \cap C \cap D$.

2.15 Consider the sample space S = {copper, sodium, nitrogen, potassium, uranium, oxygen, zinc} and the events

A = {copper, sodium, zinc},
B = {sodium, nitrogen, potassium},
C = {oxygen}.

List the elements of the sets corresponding to the following events:

(a) B ;
(b) $B \cup C$;
(c) $(A \cap B) \cup C$;
(d) $A \cap C$;
(e) $A \cap B \cap C$;

(f) $(A \cup B) \cap (A \cap C)$.

2.16 If $S = \{x \mid 0 < x < 12\}$, $M = \{x \mid 1 < x < 8\}$, and $N = \{x \mid 0 < x < 6\}$, find

(a) $M \cup N$;
(b) $M \cap N$;
(c) $M \cap N$.

2.17 Let A, B, and C be events relative to the sam ple space S. Using Venn diagrams, shade the areas representing the following events:

(a) $(A \cap B)$;
(b) $(A \cup B)$;
(c) $(A \cap C) \cup B$.

2.18 Which of the following pairs of events are mutually exclusive?

(a) A golfer scoring the lowest 18-hole round in a 72-hole tournament and losing the tournament. (b) A poker player getting a flush (all cards in the same suit) and 3 of a kind on the same 5-card hand. (c) A mother giving birth to a baby girl and a set of twin daughters on the same day.

(d) A chess player losing the last game and winning the match.

2.19 Suppose that a family is leaving on a summer vacation in their camper and that M is the event that they will experience mechanical problems, T is the event that they will receive a ticket for committing a traffic violation, and V is the event that they will ar rive at a campsite with no vacancies. Referring to the Venn diagram of Figure 2.5, state in words the events represented by the following regions:

(a) region 5;
(b) region 3;
(c) regions 1 and 2 together;
(d) regions 4 and 7 together;
(e) regions 3, 6, 7, and 8 together.

2.20 Referring to Exercise 2.19 and the Venn diagram of Figure 2.5, list the numbers of the regions that represent the following events:

(a) The family will experience no mechanical problems and will not receive a ticket for a traffic violation but will arrive at a campsite with no vacancies.

(b) The family will experience both mechanical prob lems and trouble in locating a campsite with a va cancy but will not receive a ticket for a traffic vio lation.

(c) The family will either have mechanical trouble or arrive at a campsite with no vacancies but will not receive a ticket for a traffic violation.

(d) The family will not arrive at a campsite with no vacancies.

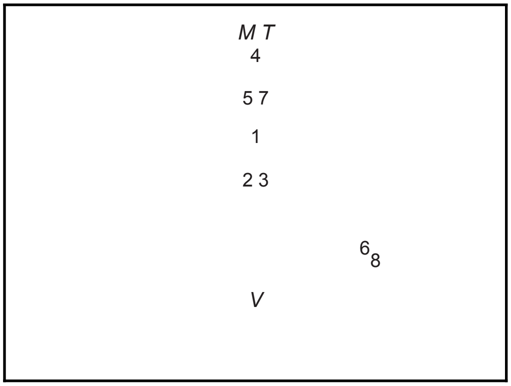

Figure 2.5: Venn diagram for Exercises 2.19 and 2.20.

2.3 Counting Sample Points

One of the problems that the statistician must consider and attempt to evaluate is the element of chance associated with the occurrence of certain events when an experiment is performed. These problems belong in the field of probability, a subject to be introduced in Section 2.4. In many cases, we shall be able to solve a probability problem by counting the number of points in the sample space without actually listing each element. The fundamental principle of counting, often referred to as the multiplication rule, is stated in Rule 2.1.

Rule 2.1:

If an operation can be performed

second operation can be performed in n_2 wa
be performed together in $n_1 n_2$ ways.

Example 2.13: How many sample points are there in the sample space when a pair of dice is thrown once?

Solution: The first die can land face-up in any one of $n_1 = 6$ ways. For each of these 6 ways, the second die can also land face-up in $n_2 = 6$ ways. Therefore, the pair of dice can land in $n_1 n_2 = (6)(6) = 36$ possible ways.

Example 2.14: A developer of a new subdivision offers prospective home buyers a choice of Tudor, rustic, colonial, and traditional exterior styling in ranch, two-story, and split-level floor plans. In how many different ways can a buyer order one of these homes?

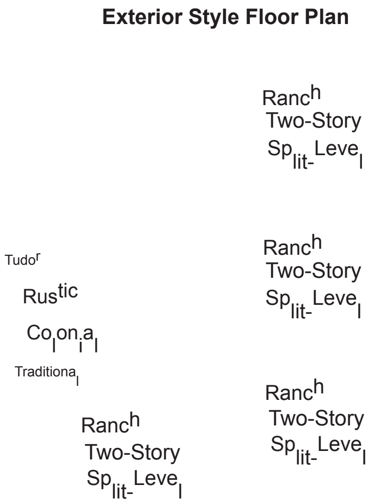

Figure 2.6: Tree diagram for Example 2.14.

Solution: Since $n_1 = 4$ and $n_2 = 3$, a buyer must choose from

$$n_1n_2 = (4)(3) = 12 \text{ possible homes.}$$

The answers to the two preceding examples can be verified by constructing tree diagrams and counting the various paths along the branches. For instance,

in Example 2.14 there will be $n_1 = 4$ branches corresponding to the different exterior styles, and then there will be $n_2 = 3$ branches extending from each of these 4 branches to represent the different floor plans. This tree diagram yields the $n_1n_2 = 12$ choices of homes given by the paths along the branches, as illustrated in Figure 2.6.

Example 2.15: If a 22-member club needs to elect a chair and a treasurer, how many different ways can these two to be elected?

Solution: For the chair position, there are 22 total possibilities. For each of those 22 pos sibilities, there are 21 possibilities to elect the treasurer. Using the multiplication rule, we obtain $n_1 \times n_2 = 22 \times 21 = 462$ different ways.

The multiplication rule, Rule 2.1 may be extended to cover any number of operations. Suppose, for instance, that a customer wishes to buy a new cell phone and can choose from $n_1 = 5$ brands, $n_2 = 5$ sets of capability, and $n_3 = 4$ colors. These three classifications result in $n_1n_2n_3 = (5)(5)(4) = 100$ different ways for a customer to order one of these phones. The generalized multiplication rule covering k operations is stated in the following.

operations can be performed in $n_1n_2 \cdots n_k$ w

Rule 2.2:

If an operation can be performed
operation can be performed in n
operation can be performed in n_3

Example 2.16: Sam is going to assemble a computer by himself. He has the choice of chips from two brands, a hard drive from four, memory from three, and an accessory bundle from five local stores. How many different ways can Sam order the parts?

Solution: Since $n_1 = 2$, $n_2 = 4$, $n_3 = 3$, and $n_4 = 5$, there are

$$n_l \times n_2 \times n_3 \times n_4 = 2 \times 4 \times 3 \times 5 = 120$$

different ways to order the parts.

Example 2.17: How many even four-digit numbers can be formed from the digits 0, 1, 2, 5, 6, and 9 if each digit can be used only once?

Solution: Since the number must be even, we have only $n_1 = 3$ choices for the units position. However, for a four-digit number the thousands position cannot be 0. Hence, we consider the units position in two parts, 0 or not 0. If the units position is 0 (i.e., $n_1 = 1$), we have $n_2 = 5$ choices for the thousands position, $n_3 = 4$ for the hundreds position, and $n_4 = 3$ for the tens position. Therefore, in this case we have a total of

$$n_1n_2n_3n_4 = (1)(5)(4)(3) = 60$$

even four-digit numbers. On the other hand, if the units position is not 0 (i.e., $n_1 =$

2), we have $n_2 = 4$ choices for the thousands position, $n_3 = 4$ for the hundreds position, and $n_4 = 3$ for the tens position. In this situation, there are a total of

$$n_1n_2n_3n_4 = (2)(4)(4)(3) = 96$$

even four-digit numbers.

Since the above two cases are mutually exclusive, the total number of even four-digit numbers can be calculated as 60 + 96 = 156.

Frequently, we are interested in a sample space that contains as elements all possible orders or arrangements of a group of objects. For example, we may want to know how many different arrangements are possible for sitting 6 people around a table, or we may ask how many different orders are possible for drawing 2 lottery tickets from a total of 20. The different arrangements are called permutations.

Definition 2.7:

A permutation is an arrangeme

Consider the three letters *a*, *b*, and *c*. The possible permutations are *abc*, *acb*, *bac*, *bca*, *cab*, and *cba*. Thus, we see that there are 6 distinct arrangements. Using Rule 2.2, we could arrive at the answer 6 without actually listing the different orders by the following arguments: There are $n_1 = 3$ choices for the first position. No matter which letter is chosen, there are always $n_2 = 2$ choices for the second position. No matter which two letters are chosen for the first two positions, there is only $n_3 = 1$ choice for the last position, giving a total of

$$n_1n_2n_3 = (3)(2)(1) = 6 \text{ permutations}$$

by Rule 2.2. In general, *n* distinct objects can be arranged in

$$n(n-1)(n-2)\cdots(3)(2)(1) \text{ ways.}$$

There is a notation for such a number.

Definition 2.8:

For any non-negative integer *n*, *n*!, callec

as $n! = n(n-1)\cdots(2)($

with special case 0! = 1.

Theorem 2.1:

Using the argument above, we arrive at the following theorem.

The number of permutations of *n* objects is

The number of permutations of the four letters *a*, *b*, *c*, and *d* will be 4! = 24. Now consider the number of permutations that are possible by taking two letters at a time from four. These would be *ab*, *ac*, *ad*, *ba*, *bc*, *bd*, *ca*, *cb*, *cd*, *da*, *db*, and *dc*. Using Rule 2.1 again, we have two positions to fill, with $n_1 = 4$ choices for the first and then $n_2 = 3$ choices for the second, for a total of

$n_1 n_2 = (4)(3) = 12$

permutations. In general, n distinct objects taken

r at a time can be arranged in $n(n - 1)(n -$

$2)\cdots(n - r + 1)$

ways. We represent this product by the symbol $_nP_r = n!$

$(n - r)!\cdot$

As a result, we have the theorem that follows.

Theorem 2.2:

The number of permutations o

is ,

$(n - r$

Example 2.18: In one year, three awards (research, teaching, and service) will be given to a class of 25 graduate students in a statistics department. If each student can receive at most one award, how many possible selections are there?

Solution: Since the awards are distinguishable, it is a permutation problem. The total number of sample points is

$(25 - 3)! = 25!$

800.

$_{25}P_3 = 25!$

$22! = (25)(24)(23) = 13,$

Example 2.19: A president and a treasurer are to be chosen from a student club consisting of 50 people. How many different choices of officers are possible if

(a) there are no restrictions;

(b) *A* will serve only if he is president;

(c) *B* and *C* will serve together or not at all;

(d) *D* and *E* will not serve together?

Solution: (a) The total number of choices of officers, without any restrictions, is

$_{50}P_2 = 50!$

$48! = (50)(49) = 2450.$

(b) Since *A* will serve only if he is president, we have two situations here: (i) *A* is selected as the president, which yields 49 possible outcomes for the treasurer's position, or (ii) officers are selected from the remaining 49 people without *A*, which has the number of choices $_{49}P_2 = (49)(48) = 2352$. Therefore, the total number of choices is 49 + 2352 = 2401.

(c) The number of selections when *B* and *C* serve together is 2. The number of selections when both *B* and *C* are not chosen is $_{48}P_2 = 2256$. Therefore, the total number of choices in this situation is 2 + 2256 = 2258.

(d) The number of selections when *D* serves as an officer but not *E* is (2)(48) = 96, where 2 is the number of positions *D* can take and 48 is the number of selections of the other officer from the remaining people in the club except *E*. The number of selections when *E* serves as an officer but not *D* is also (2)(48) = 96. The number of selections when both *D* and *E* are not chosen is $_{48}P_2 = 2256$. Therefore, the total number of choices is (2)(96) + 2256 = 2448. This problem also has another short solution: Since *D* and *E* can only serve together in 2 ways, the answer is 2450 − 2 = 2448.

Permutations that occur by arranging objects in a circle are called circular permutations. Two circular permutations are not considered different unless corresponding objects in the two arrangements are preceded or followed by a dif ferent object as we proceed in a clockwise direction. For example, if 4 people are playing bridge, we do not have a new permutation if they all move one position in a clockwise direction. By considering one person in a fixed position and arranging the other three in 3! ways, we find that there are 6 distinct arrangements for the bridge game.

Theorem 2.3: Theorem 2.4:

The number of permutations of

Obviously, if the letters *b* and *c* are both equal to *x*, then the 6 permutations of the letters *a*, *b*, and *c* become *axx*, *axx*, *xax*, *xax*, *xxa*, and *xxa*, of which only 3 are distinct. Therefore, with 3 letters, 2 being the same, we have 3!/2! = 3 distinct permutations. With 4 different letters *a*, *b*, *c*, and *d*, we have 24 distinct permutations. If we let *a* = *b* = *x* and *c* = *d* = *y*, we can list only the following distinct permutations: *xxyy*, *xyxy*, *yxxy*, *yyxx*, *xyyx*, and *yxyx*. Thus, we have 4!/(2! 2!) = 6 distinct permutations.

The number of distinct permutations of *n* t
n_2 of a second kind, ... , n_k of a *k*th kind is

$n!$

$n_1!n_2! \cdots r$

So far we have considered permutations of distinct objects. That is, all the objects were completely different or distinguishable.

Example 2.20: In a college football training session, the defensive coordinator needs to have 10 players standing in a row. Among these 10 players, there are 1 freshman, 2 sopho mores, 4 juniors, and 3 seniors. How many different ways can they be arranged in a row if only their class level will be distinguished?

Solution: Directly using Theorem 2.4, we find that the total number of arrangements is

$$\frac{10!}{1!\ 2!\ 4!\ 3!} = 12,600.$$

Often we are concerned with the number of ways of partitioning a set of *n* objects into *r* subsets called cells. A partition has been achieved if the intersection of every possible pair of the *r* subsets is the empty set *φ* and if the union of all subsets gives the original set. The order of the elements within a cell is of no importance. Consider the set *{a, e, i, o, u}*. The possible partitions into two cells in which the first cell contains 4 elements and the second cell 1 element are

{(a, e, i, o),(u)}, {(a, i, o, u),(e)}, {(e, i, o, u),(a)}, {(a, e, o, u),(i)}, {(a, e, i, u),(o)}.

We see that there are 5 ways to partition a set of 4 elements into two subsets, or cells, containing 4 elements in the first cell and 1 element in the second.

The number of partitions for this illustration is denoted by the symbol

$$\binom{5}{4,\ 1} = \frac{5!}{4!\ 1!} = 5,$$

where the top number represents the total number of elements and the bottom numbers represent the number of elements going into each cell. We state this more generally in Theorem 2.5.

Theorem 2.5:

The number of ways of partitioning a set
elements in the first cell, n_2 elements in t

n

where $n_1 + n_2 + \cdots + n_r = n$.

n_1 , n_2 ,..

Example 2.21: In how many ways can 7 graduate students be assigned to 1 triple and 2 double hotel rooms during a conference?

Solution: The total number of possible partitions would be

7

3, 2, 2 = 7! 3! 2! 2! = 210.

In many problems, we are interested in the number of ways of selecting r objects from n without regard to order. These selections are called combinations. A combination is actually a partition with two cells, the one cell containing the r objects selected and the other cell containing the ($n-r$) objects that are left. The number of such combinations, denoted by

n

$r, n - r$, is usually shortened to n r ,

n

= n!

r!(n −

r

Theorem 2.6:

since the number of elements in the second cell must be $n - r$.

The number of combinations o

Example 2.22: A young boy asks his mother to get 5 Game-Boy^{TM} cartridges from his collection of 10 arcade and 5 sports games. How many ways are there that his mother can get 3 arcade and 2 sports games?

Solution: The number of ways of selecting 3 cartridges from 10 is

10 3

3! (10 − 3)! = 120.

= 10!

The number of ways of selecting 2 cartridges from 5 is

5 2 = 5! 2! 3! = 10.

Exercises 71

Using the multiplication rule (Rule 2.1) with $n_1 = 120$ and $n_2 = 10$, we have (120)(10) = 1200 ways.

Example 2.23: How many different letter arrangements can be made from the letters in the word *STATISTICS*?

Solution: Using the same argument as in the discussion for Theorem 2.6, in this example we can actually apply Theorem 2.5 to obtain

10 3! 3! 2! 1! 1! = 50,
3, 3, 2, 1, 1 400.

= 10!

Here we have 10 total letters, with 2 letters (S, T) appearing 3 times each, letter I appearing twice, and letters A and C appearing once each. On the other hand, this result can be directly obtained by using Theorem 2.4.

Exercises

2.21 Registrants at a large convention are offered 5 sightseeing tours on each of 3 days. In how many ways can a person arrange to go on a sightseeing tour planned by this convention?

2.22 In a medical study, patients are classified in 8 ways according to whether they have blood type AB^+, AB^-, A^+, A^-, B^+, B^-, O^+, or O^-, and also accord ing to whether their blood pressure is low, normal, or high. Find the number of ways in which a patient can be classified.

2.23 If an experiment consists of tossing a coin and then drawing a letter at random from the English alphabet, how many points are there in the sample space?

2.24 Students at a private liberal arts college are classified as being freshmen, sophomores, juniors, or seniors, and also according to whether they are male or female. Find the total number of possible classifications for the students of that college.

2.25 A certain brand of shoes comes in 6 different styles, with each style available in 5 distinct colors. If the store wishes to display pairs of these shoes showing all of its various styles and colors, how many different pairs will the store have on display?

2.26 A California study concluded that following 7 simple health rules can extend a man's life by 11 years on the average and a woman's life by 7 years. These 7 rules are as follows: no smoking, get regular exer cise, use alcohol only in moderation, get 7 to 8 hours of sleep, maintain proper weight, eat breakfast, and do not eat between meals. In how many ways can a person adopt 5 of these rules to follow

(a) if the person presently violates all 7 rules? (b) if the person never drinks and always eats break fast?

2.27 A developer of a new subdivision offers a prospective home buyer a choice of 3 designs, 3 different heating systems, a garage or carport, and a patio or screened porch. How many different plans are available to this buyer?

2.28 A drug for the relief of asthma can be purchased from 4 different manufacturers in liquid, tablet, or capsule form, all of which come in regular and extra strength. How many different ways can a doctor prescribe the drug for a patient suffering from asthma?

2.29 In a fuel economy study, 3 race cars are tested using 6 different brands of gasoline at 7 test sites located in different regions of the country. If 2 drivers are used in the study, and test runs are made once under each distinct set of conditions, how many test runs are needed?

2.30 In how many different ways can a true-false test consisting of 10 questions be answered?

2.31 A witness to a hit-and-run accident told the po lice that the license number contained the letters RLH followed by 4 digits, the first of which was a 5. If the witness cannot recall the last 3 digits, but is cer tain that all 4 digits are different, find the maximum number of automobile registrations that the police may have to check.

2.32 (a) In how many ways can 6 people be lined up to get on a bus?

(b) If 3 specific persons, among 6, insist on following each other, how many ways are possible? (c) If 2 specific persons, among 6, refuse to follow each other, how many ways are possible?

2.33 If a multiple-choice test consists of 6 questions, each with 4 possible answers of which only 1 is correct,

(a) in how many different ways can a student check off one answer to each question?

(b) in how many ways can a student check off one answer to each question and get all the answers wrong?

2.34 (a) How many distinct permutations can be made from the letters of the word *COLUMNS*? (b) How many of these permutations start with the let ter M?

2.35 A contractor wishes to build 8 houses, each different in design. In how many ways can they place these houses on a street if 5 lots are on one side of the street and 3 lots are on the opposite side?

2.36 (a) How many three-digit numbers can be formed from the digits 0, 1, 2, 3, 4, 5, and 6 if each digit can be used only once?

(b) How many of these are odd numbers?

(c) How many are greater than 330?

2.37 In how many ways can 3 boys and 4 girls sit in a row if the boys and girls must alternate?

2.38 Three married couples have bought 6 seats in the same row for a concert. In how many different ways can they be seated

(a) with no restrictions?

(b) if each couple is to sit together?

2.4 Probability of an Event

(c) if all the men sit together to the right of all the women?

2.39 In a regional spelling bee, the 8 finalists consist of 3 boys and 5 girls. Find the number of sample points in the sample space *S* for the number of possible orders at the conclusion of the contest for

(a) all 8 finalists;

(b) the first 3 positions.

2.40 In how many ways can 6 starting positions on a basketball team be filled with 9 men who can play any of the positions?

2.41 Find the number of ways that 7 teachers can be assigned to 4 sections of an introductory psychology course if no teacher is assigned to more than one section.

2.42 Three lottery tickets for the first, second, and third prizes are drawn from a group of 20 tickets. Find the number of sample points, in *S*, for awarding the 3 prizes if each contestant holds only 1 ticket.

2.43 In how many ways can 6 different trees be planted in a circle?

2.44 In how many ways can a caravan of 9 covered wagons from Arizona be arranged in a circle?

2.45 How many distinct permutations can be made from the letters of the word *INFINITY* ?

2.46 In how many ways can 2 oaks, 5 pines, and 3 maples be arranged along a property line if one does not distinguish among trees of the same kind?

2.47 How many ways are there to select 3 candidates from 7 equally qualified recent graduates for openings in an accounting firm?

2.48 How many ways are there that no two students will have the same birth date in a class of size 50?

Perhaps it was humankind's unquenchable thirst for gambling that led to the early development of probability theory. In an effort to increase their winnings, gamblers called upon mathematicians to provide optimum strategies for various games of chance. Some of the mathematicians providing these strategies were Pascal, Leibniz, Fermat, and James Bernoulli. As a result of this development of probability theory, statistical inference, with all its predictions and generalizations, has branched out far beyond games of chance to encompass many other fields associated with chance occurrences, such as politics, business, weather forecasting,

and scientific research. For these predictions and generalizations to be reasonably accurate, an understanding of basic probability theory is essential.

What do we mean when we make the statement "John will probably win the tennis match," or "I have a fifty-fifty chance of getting an even number when a die is tossed," or "The university is not likely to win the football game tonight," or "Most of our graduating class will likely be married within 3 years"? In each case, we are expressing an outcome of which we are not certain, but owing to past information or from an understanding of the structure of the experiment, we have some degree of confidence in the validity of the statement.

Throughout the remainder of this chapter, we consider only those experiments for which the sample space contains a finite number of elements. The likelihood of the occurrence of an event resulting from such a statistical experiment is evaluated by means of a set of real numbers, called weights or probabilities, ranging from 0 to 1. To every point in the sample space we assign a probability such that the sum of all probabilities is 1. If we have reason to believe that a certain sample point is quite likely to occur when the experiment is conducted, the probability assigned should be close to 1. On the other hand, a probability closer to 0 is assigned to a sample point that is not likely to occur. In many experiments, such as tossing a coin or a die, all the sample points have the same chance of occurring and are assigned equal probabilities. For points outside the sample space, that is, for simple events that cannot possibly occur, we assign a probability of 0.

To find the probability of an event *A*, we sum all the probabilities assigned to the sample points in *A*. This sum is called the probability of *A* and is denoted by *P*(*A*).

then $P(A_1 \cup A_2 \cup A_3 \cup \cdots) = P(A$

Definition 2.9:

The probability of an event *A* is

A. Therefore,

$0 \leq P(A) \leq 1,\ P($

Furthermore, if $A_1, A_2, A_3, ..$

Example 2.24: A coin is tossed twice. What is the probability that at least 1 head occurs?

Solution: The sample space for this experiment is

$$S = \{HH, HT, TH, TT\}.$$

If the coin is balanced, each of these outcomes is equally likely to occur. Therefore, we assign a probability of ω to each sample point. Then $4\omega = 1$, or $\omega = 1/4$. If *A* represents the event of at least 1 head occurring, then

$$A = \{HH, HT, TH\} \text{ and } P(A) = \frac{1}{4} + \frac{1}{4} + \frac{1}{4} = \frac{3}{4}\cdot$$

Example 2.25: A die is loaded in such a way that an even number is twice as likely to occur as an odd number. If *E* is the event that a number less than 4 occurs on a single toss of the die, find *P*(*E*).

Solution: The sample space is $S = \{1, 2, 3, 4, 5, 6\}$. We assign a probability of *w* to each odd number and a probability of 2*w* to each even number. Since the sum of the probabilities must be 1, we have $9w = 1$ or $w = 1/9$. Hence, probabilities of 1/9 and 2/9 are assigned to each odd and even number, respectively. Therefore,

$$E = \{1, 2, 3\} \text{ and } P(E) = \frac{1}{9} + \frac{2}{9} + \frac{1}{9} = \frac{4}{9}\cdot$$

Example 2.26: In Example 2.25, let *A* be the event that an even number turns up and let *B* be the event that a number divisible by 3 occurs. Find $P(A \cup B)$ and $P(A \cap B)$. *Solution*: For the events $A = \{2, 4, 6\}$ and $B = \{3, 6\}$, we have

$$A \cup B = \{2, 3, 4, 6\} \text{ and } A \cap B = \{6\}.$$

By assigning a probability of 1/9 to each odd number and 2/9 to each even number, we have

$$P(A \cup B) = \frac{2}{9} + \frac{1}{9} + \frac{2}{9} + \frac{2}{9} = \frac{7}{9} \text{and } P(A \cap B) = \frac{2}{9}\cdot$$

If the sample space for an experiment contains N elements, all of which are equally likely to occur, we assign a probability equal to $1/N$ to each of the N points. The probability of any event A containing n of these N sample points is then the ratio of the number of elements in A to the number of elements in S.

Rule 2.3:

> If an experiment can result in an
> and if exactly n of these outcome
> of event A is
>
> P

Example 2.27: A statistics class for engineers consists of 25 industrial, 10 mechanical, 10 electrical, and 8 civil engineering students. If a person is randomly selected by the instruc tor to answer a question, find the probability that the student chosen is (a) an industrial engineering major and (b) a civil engineering or an electrical engineering major.

Solution: Denote by I, M, E, and C the students majoring in industrial, mechanical, electrical, and civil engineering, respectively. The total number of students in the class is 53, all of whom are equally likely to be selected.

(a) Since 25 of the 53 students are majoring in industrial engineering, the probability of event I, selecting an industrial engineering major at random, is

$$P(I) = \frac{25}{53}\cdot$$

(b) Since 18 of the 53 students are civil or electrical engineering majors, it follows that

$$P(C \cup E) = \frac{18}{53}\cdot$$

Example 2.28: In a poker hand consisting of 5 cards, find the probability of holding 2 aces and 3 jacks.

Solution: The number of ways of being dealt 2 aces from 4 cards is

$$\binom{4}{2} = \frac{4!}{2!\,2!} = 6,$$

and the number of ways of being dealt 3 jacks from 4 cards is

$$\binom{4}{3} = \frac{4!}{3!\,1!} = 4.$$

By the multiplication rule (Rule 2.1), there are $n = (6)(4) = 24$ hands with 2 aces and 3 jacks. The total number of 5-card poker hands, all of which are equally likely, is

$$N = \binom{52}{5} = \frac{52!}{5!\,47!} = 2{,}598{,}960.$$

Therefore, the probability of getting 2 aces and 3 jacks in a 5-card poker hand is $P(C) = 24$

$2, 598, 960 = 0.9 \times 10^{-5}$.

If the outcomes of an experiment are not equally likely to occur, the probabilities must be assigned on the basis of prior knowledge or experimental evidence. For example, if a coin is not balanced, we could estimate the probabilities of heads and tails by tossing the coin a large number of times and recording the outcomes. According to the relative frequency definition of probability, the true probabilities would be the fractions of heads and tails that occur in the long run. Another intuitive way of understanding probability is the indifference approach. For instance, if you have a die that you believe is balanced, then using this indifference approach, you determine that the probability that each of the six sides will show up after a throw is 1/6.

To find a numerical value that represents adequately the probability of winning at tennis, we must depend on our past performance at the game as well as that of the opponent and, to some extent, our belief in our ability to win. Similarly, to find the probability that a horse will win a race, we must arrive at a probability based on the previous records of all the horses entered in the race as well as the records of the jockeys riding the horses. Intuition would undoubtedly also play a part in determining the size of the bet that we might be willing to wager. The use of intuition, personal beliefs, and other indirect information in arriving at probabilities is referred to as the subjective definition of probability.

In most of the applications of probability in this book, the relative frequency interpretation of probability is the operative one. Its foundation is the statistical experiment rather than subjectivity, and it is best viewed as the limiting relative frequency. As a result, many applications of probability in science and engineering must be based on experiments that can be repeated. Less objective notions of probability are encountered when we assign probabilities based on prior information and opinions, as in “There is a good chance that the Giants will lose the Super Bowl.” When opinions and prior information differ from individual to individual, subjective probability becomes the relevant resource. In Bayesian statistics (see Chapter 18), a more subjective interpretation of probability will be used, based on an elicitation of prior probability information.

2.5 Additive Rules

Often it is easiest to calculate the probability of some event from known probabilities of other events. This may well be true if the event in question can be represented as the union of two other events or as the complement of some event. Several important laws that frequently simplify the computation of probabilities follow. The first, called the additive rule, applies to unions of events.

S

Theorem 2.7:

If *A* and *B* are two events, then

A B A B

$P(A \cup B) =$

Figure 2.7: Additive rule of probability.

Proof : Consider the Venn diagram in Figure 2.7. The $P(A \cup B)$ is the sum of the probabilities of the sample points in $A \cup B$. Now $P(A) + P(B)$ is the sum of all the probabilities in A plus the sum of all the probabilities in B. Therefore, we have added the probabilities in $(A \cap B)$ twice. Since these probabilities add up to $P(A \cap B)$, we must subtract this probability once to obtain the sum of the probabilities in $A \cup B$.

Corollary 2.1:

If A and B are mutually exclusi

$P(A \cup$

Corollary 2.1 is an immediate result of Theorem 2.7, since if A and B are mutually exclusive, $A \cap B = 0$ and then $P(A \cap B) = P(\varphi) = 0$. In general, we can write Corollary 2.2.

A collection of events $\{A_1, A_2, ..., A_n\}$ of a sample space S is called a partition of S if $A_1, A_2, ..., A_n$ are mutually exclusive and $A_1 \cup A_2 \cup \cdots \cup A_n = S$. Thus, we have

Corollary 2.2:

If $A_1, A_2, ..., A_n$ is a partition of sample spa

$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2$

Corollary 2.3: Theorem 2.8:

As one might expect, Theorem 2.7 extends in an analogous fashion.

For three events A, B, and C,

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

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

If $A_1, A_2, ..., A_n$ are mutually exc

$P(A_1 \cup A_2 \cup \cdots \cup A_n$

Example 2.29: John is going to graduate from an industrial engineering department in a university by the end of the semester. After being interviewed at two companies he likes, he assesses that his probability of getting an offer from company A is 0.8, and his probability of getting an offer from company B is 0.6. If he believes that the probability that he will get offers from both companies is 0.5, what is the probability that he will get at least one offer from these two companies?

Solution: Using the additive rule, we have

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)=0.8+0.6-0.5=0.9.$$

Example 2.30: What is the probability of getting a total of 7 or 11 when a pair of fair dice is tossed?

Solution: Let A be the event that 7 occurs and B the event that 11 comes up. Now, a total of 7 occurs for 6 of the 36 sample points, and a total of 11 occurs for only 2 of the sample points. Since all sample points are equally likely, we have $P(A) = 1/6$ and $P(B) = 1/18$. The events A and B are mutually exclusive, since a total of 7 and 11 cannot both occur on the same toss. Therefore,

$$P(A \cup B) = P(A) + P(B) = \frac{1}{6} + \frac{1}{18} = \frac{2}{9}\cdot$$

This result could also have been obtained by counting the total number of points for the event $A \cup B$, namely 8, and writing

$$P(A \cup B) = \frac{n}{N} = \frac{8}{36} = \frac{2}{9}\cdot$$

Theorem 2.7 and its three corollaries should help the reader gain more insight into probability and its interpretation. Corollaries 2.1 and 2.2 suggest the very intuitive result dealing with the probability of occurrence of at least one of a number of events, no two of which can occur simultaneously. The probability that at least one occurs is the sum of the probabilities of occurrence of the individual events. The third corollary simply states that the highest value of a probability (unity) is assigned to the entire sample space S.

Example 2.31: If the probabilities are, respectively, 0.09, 0.15, 0.21, and 0.23 that a person purchasing a new automobile will choose the color green, white, red, or blue, what is the probability that a given buyer will purchase a new automobile that comes in one of those colors?

Solution: Let G, W, R, and B be the events that a buyer selects, respectively, a green, white, red, or blue automobile. Since these four events are mutually exclusive, the probability is

$$\begin{aligned} P(G \cup W \cup R \cup B) &= P(G) + P(W) + P(R) + P(B) \\ &= 0.09 + 0.15 + 0.21 + 0.23 = 0.68. \end{aligned}$$

Often it is more difficult to calculate the probability that an event occurs than it is to calculate the probability that the event does not occur. Should this be the case for some event A, we simply find $P(A\ )$ first and then, using Theorem 2.7, find $P(A)$ by subtraction.

$P(A) + P(A\ )$

Theorem 2.9:

If A and A are complementary e

Proof : Since $A \cup A\ = S$ and the sets A and A are disjoint,

$$1 = P(S) = P(A \cup A\ ) = P(A) + P(A\ ).$$

Example 2.32: If the probabilities that an automobile mechanic will service 3, 4, 5, 6, 7, or 8 or more cars on any given workday are, respectively, 0.12, 0.19, 0.28, 0.24, 0.10, and 0.07, what is the probability that he will service at least 5 cars on his next day at work?

Solution: Let E be the event that at least 5 cars are serviced. Now, $P(E)=1-P(E\ )$, where E is the event that fewer than 5 cars are serviced. Since

$$P(E\ )=0.12+0.19=0.31,$$

it follows from Theorem 2.9 that

$$P(E)=1-0.31=0.69.$$

Example 2.33: Suppose the manufacturer's specifications for the length of a certain type of computer cable are 2000 ± 10 millimeters. In this industry, it is known that small cable is just as likely to be defective (not meeting specifications) as large cable. That is,

the probability of randomly producing a cable with length exceeding 2010 millimeters is equal to the probability of producing a cable with length smaller than 1990 millimeters. The probability that the production procedure meets specifications is known to be 0.99.

(a) What is the probability that a cable selected randomly is too large?

(b) What is the probability that a randomly selected cable is larger than 1990 millimeters?

Solution: Let M be the event that a cable meets specifications. Let S and L be the events that the cable is too small and too large, respectively. Then

(a) $P(M)=0.99$ and $P(S)=P(L)=(1-0.99)/2=0.005$.

(b) Denoting by X the length of a randomly selected cable, we have

$$P(1990\leq X\leq 2010)=P(M)=0.99.$$

Since $P(X\geq 2010)=P(L)=0.005$,

$$P(X\geq 1990)=P(M)+P(L)=0.995.$$

This also can be solved by using Theorem 2.9:

$$P(X\geq 1990)+P(X<1990)=1.$$

Thus, $P(X\geq 1990)=1-P(S)=1-0.005=0.995$.

Exercises

2.49 Find the errors in each of the following statements:

(a) The probabilities that an automobile salesperson will sell 0, 1, 2, or 3 cars on any given day in February are, respectively, 0.19, 0.38, 0.29, and 0.15.

(b) The probability that it will rain tomorrow is 0.40, and the probability that it will not rain tomorrow is 0.52.

(c) The probabilities that a printer will make 0, 1, 2, 3, or 4 or more mistakes in setting a document are, respectively, 0.19, 0.34, −0.25, 0.43, and 0.29.

(d) On a single draw from a deck of playing cards, the probability of selecting a heart is 1/4, the probability of selecting a black card is 1/2, and the probability of selecting both a heart and a black card is 1/8.

2.50 Assuming that all elements of S in Exercise 2.8 on page 62 are equally likely to occur, find (a) the probability of event A;

(b) the probability of event C;

(c) the probability of event $A\cap C$.

2.51 A box contains 500 envelopes, of which 50 contain \$100 in cash, 150 contain \$25, and 300 contain

$10. An envelope may be purchased for $25. What is the sample space for the different amounts of money? Assign probabilities to the sample points and then find the probability that the first envelope purchased con tains less than $100.

2.52 Suppose that in a senior college class of 500 stu dents it is found that 210 smoke, 258 drink alcoholic beverages, 216 eat between meals, 122 smoke and drink alcoholic beverages, 83 eat between meals and drink alcoholic beverages, 97 smoke and eat between meals, and 52 engage in all three of these bad health practices. If a member of this senior class is selected at random, find the probability that the student

(a) smokes but does not drink alcoholic beverages; (b) eats between meals and drinks alcoholic beverages but does not smoke;

(c) neither smokes nor eats between meals.

2.53 TheprobabilitythatanAmericanindustrywillbe located in Shanghai, China, is 0.6, the probability that