

---

Anthony Almudevar, PhD

© 2014 – 2020

# INTRODUCTION TO STATISTICAL METHODOLOGY

A First Course in Statistics

# Contents

<b>1</b>	<b>Introduction</b>	<b>8</b>
<b>I</b>	<b>Probability</b>	<b>11</b>
<b>2</b>	<b>Probability - Basic Theory</b>	<b>12</b>
2.1	Definitions . . . . .	13
2.2	Set Algebra . . . . .	13
2.3	Rules of Probabilities . . . . .	15
2.4	Venn Diagram . . . . .	17
2.5	Independence . . . . .	18
2.6	Conditional Probability . . . . .	20
2.7	The Law of Total Probability . . . . .	22
<b>3</b>	<b>Combinatorics and Equiprobable Distributions</b>	<b>24</b>
3.1	Counting Rules . . . . .	24
<b>4</b>	<b>Random Variables</b>	<b>33</b>
4.1	Types of Sample Spaces . . . . .	34
4.2	Probability Mass Functions and Densities . . . . .	34
4.3	Cumulative Distribution Functions . . . . .	37
4.4	Quantiles and Percentiles . . . . .	42
4.5	Expected Value . . . . .	45
4.6	Functions of Random Variables . . . . .	48
4.7	Variance . . . . .	51
4.8	Random Variables and Independence . . . . .	53
4.9	Sums and Averages of Random Variables . . . . .	57
4.10	Some Common Discrete Random Variables . . . . .	60
4.11	Some Common Continuous Random Variables . . . . .	67
4.12	The Normal Distribution . . . . .	69
4.13	Central Limit Theorem . . . . .	75
4.14	Normal Approximation to the Binomial . . . . .	77
4.15	The $\chi^2$ , $t$ - and $F$ - distributions . . . . .	81
4.16	Survival Times . . . . .	83
4.17	Power Law Distributions . . . . .	84

4.18 Random Variables in R . . . . .	88
<b>5 Bayes' Theorem and Classification</b>	<b>90</b>
5.1 Odds . . . . .	92
5.2 The Bayesian Model . . . . .	93
5.3 The Fallacy of the Transposed Conditional . . . . .	95
5.4 Diagnostic Testing - Basic Definitions . . . . .	96
5.5 The Odds Ratio . . . . .	101
<b>6 Simulation of Random Variables</b>	<b>102</b>
6.1 Linear Congruental Generators . . . . .	103
6.2 The Inverse Transformation Method . . . . .	107
6.3 Simulation of Discrete Random Variables . . . . .	108
<b>7 Stochastic Processes</b>	<b>110</b>
7.1 Poisson Process . . . . .	110
7.2 Markov Chains . . . . .	112
7.3 Birth and Death Processes . . . . .	123
7.4 Queueing Systems . . . . .	125
<b>II Statistics - Introduction</b>	<b>129</b>
<b>8 Statistical Summaries of Central Tendency</b>	<b>130</b>
8.1 The Role of Variation in Statistics . . . . .	131
8.2 The Sample Mean . . . . .	133
8.3 Order Statistics . . . . .	134
8.4 The Median . . . . .	135
8.5 The Trimmed Mean . . . . .	136
8.6 Sample Quantiles and Percentiles . . . . .	137
8.7 Random Samples and Data Homogeneity . . . . .	139
<b>9 Graphical Summaries</b>	<b>141</b>
9.1 Stem and Leaf Plots . . . . .	141
9.2 Histograms . . . . .	143
9.3 Boxplots . . . . .	146
<b>10 Properties of Data</b>	<b>153</b>
10.1 Types of Data . . . . .	153
10.2 Distributions . . . . .	154
10.3 Central Tendency and Variability . . . . .	155
10.4 Coefficient of Variation . . . . .	163
10.5 Symmetry and Skewness . . . . .	163
10.6 The Empirical Rule . . . . .	165
10.7 Quantile Plots . . . . .	167
10.8 Transformations . . . . .	169

<b>11 Relationships Between Variables</b>	<b>180</b>
11.1 Relationships Between Categorical Variables . . . . .	181
11.2 Relationships Between One Categorical Variable and One Numerical Variable . . . . .	184
11.3 Relationships Between Numerical Variables . . . . .	185
11.4 Scatter Plots in R . . . . .	191
<b>III Statistics - Inference</b>	<b>194</b>
<b>12 Confidence Intervals and Hypothesis Tests - Population Mean</b>	<b>195</b>
12.1 Confidence Intervals . . . . .	195
12.2 Hypothesis Tests . . . . .	197
12.3 The <i>t</i> -distribution . . . . .	200
12.4 Assumptions . . . . .	205
<b>13 Some General Definitions for Hypothesis Testing</b>	<b>207</b>
13.1 Hypothesis Testing Based on Rejection Regions . . . . .	210
13.2 Type I and Type II Errors . . . . .	213
13.3 Power . . . . .	213
13.4 Precise Definition of the Observed Level of Significance . . . . .	215
<b>14 Inference for Differences of Means</b>	<b>216</b>
14.1 Independent Samples . . . . .	217
14.2 Paired Samples . . . . .	227
14.3 The <i>t</i> -test in R . . . . .	231
14.4 Assumptions . . . . .	233
<b>15 Inference for Population Proportions</b>	<b>234</b>
15.1 Single Population Proportion . . . . .	234
15.2 Difference Between Two Population Proportions . . . . .	237
15.3 Binomial Continuity Correction . . . . .	241
15.4 Inference for the Odds Ratio . . . . .	241
15.5 Testing for Proportions in R . . . . .	243
15.6 Assumptions . . . . .	245
<b>16 Sample Size Estimation for Confidence Intervals</b>	<b>246</b>
16.1 General Approach to Sample Size Calculations: Normal Approximations . . . . .	246
16.2 Sample Size for a Confidence Interval for a Population Proportion . . . . .	248
16.3 Sample Size for a Confidence Interval for Differences in Means	249
<b>17 Power curves</b>	<b>251</b>
17.1 Power Analysis and the Noncentral <i>t</i> -distribution . . . . .	252
17.2 Sample Proportion . . . . .	254

<b>18 Inference for Variances</b>	<b>258</b>
18.1 Inference for a Single Variance . . . . .	258
18.2 Upper Confidence Bounds . . . . .	260
18.3 Sample Size Estimation . . . . .	261
18.4 IQR as Estimate of Variance . . . . .	262
18.5 Hypothesis Tests for Two Population Variances . . . . .	263
18.6 Assumptions . . . . .	265
<b>19 Nonparametric Inference</b>	<b>266</b>
19.1 Sign Test . . . . .	266
19.2 Signed Rank Test . . . . .	268
19.3 Dealing with Ties in Rank-Based Procedures . . . . .	270
19.4 Wilcoxon Rank Sum Test . . . . .	272
19.5 Assumptions . . . . .	278
<b>20 Nonparametric Inference in R</b>	<b>280</b>
20.1 Sign Test . . . . .	280
20.2 Signed Rank and Rank Sum Procedures . . . . .	282
<b>21 Inference for Correlation</b>	<b>285</b>
21.1 Inference for Correlations . . . . .	285
21.2 Sample Size Analysis for Hypothesis Tests . . . . .	287
21.3 Inference for the Pearson Correlation Coefficient When $\rho \neq 0$ . . . . .	287
21.4 Nonparametric Correlation Coefficients . . . . .	288
21.5 Inference for Correlations in R . . . . .	289
<b>22 Goodness of Fit Tests and Contingency Tables</b>	<b>294</b>
22.1 Yates's Correction . . . . .	296
22.2 Assumptions . . . . .	296
22.3 Hypothesis Tests for Contingency Tables . . . . .	296
22.4 Yates's Correction . . . . .	301
22.5 $\chi^2$ Tests in R . . . . .	301
22.6 Assumptions . . . . .	302
<b>23 ANOVA</b>	<b>303</b>
23.1 Methodology . . . . .	303
23.2 ANOVA Table . . . . .	306
23.3 Bonferroni Correction for Multiple Comparisons . . . . .	307
23.4 <i>Post hoc</i> Analysis in ANOVA . . . . .	309
23.5 Nonparametric ANOVA . . . . .	310
23.6 Assumptions . . . . .	310
<b>24 ANOVA in R</b>	<b>312</b>
24.1 Equality of Variances . . . . .	314
24.2 The Kruskal-Wallis Test for Nonparametric ANOVA . . . . .	316

<b>25 Linear Regression I</b>	<b>320</b>
25.1 Residuals . . . . .	323
25.2 ANOVA approach . . . . .	324
25.3 The Relationship Between Linear Regression and Correlation	327
25.4 Assumptions . . . . .	328
<b>26 Linear Regression II</b>	<b>329</b>
26.1 Inference of Regression Parameters . . . . .	330
26.2 Multiple Linear Regression . . . . .	336
<b>27 Linear Regression III</b>	<b>342</b>
27.1 Statistical Models . . . . .	342
27.2 ANOVA as a Model in R . . . . .	343
27.3 Linear Regression in R . . . . .	347
27.4 ANOVA and Linear Regression . . . . .	350
27.5 Residuals and <code>lm()</code> . . . . .	353
27.6 Interaction Terms . . . . .	353
27.7 Polynomial Regression . . . . .	359
<b>28 Classification and the Receiver Operator Characteristic (ROC) Curve</b>	<b>363</b>
28.1 Classifiers Based on a Numerical Risk Score . . . . .	364
28.2 ROC Curves . . . . .	370
<b>29 Simulation Methods</b>	<b>374</b>
29.1 Permutation Test . . . . .	374
29.2 The Bootstrap Procedure . . . . .	377
<b>A Distribution Tables</b>	<b>381</b>
<b>B Mathematical Review</b>	<b>409</b>
B.1 Conventions and Notation . . . . .	409
B.2 Infimum, Supremum, Minimum and Maximum . . . . .	411
B.3 Limits . . . . .	412
B.4 Matrices . . . . .	412
B.5 Geometric Series . . . . .	413
B.6 Binomial Theorem . . . . .	413
B.7 Gamma Function . . . . .	414
<b>C Introduction to R</b>	<b>415</b>
C.1 Mathematical Operations on Scalars and Vectors . . . . .	415
C.2 Data Structures in R . . . . .	431
C.3 Labels for Data Structures . . . . .	443
C.4 Programming and Functions . . . . .	448
C.5 Vectorized Calculations . . . . .	453
C.6 File Input and Output . . . . .	454

C.7 Packages . . . . .	454
C.8 Objects and Classes in R . . . . .	457
<b>D Methodological Summary . . . . .</b>	<b>465</b>
D.1 Diagnostic Testing . . . . .	465
D.2 Odds Ratios . . . . .	468
D.3 Binomial Continuity Correction . . . . .	468
D.4 Single Population Mean . . . . .	468
D.5 Difference in Population Means . . . . .	470
D.6 Inference for Population Proportions . . . . .	471
D.7 Inference for Two Population Proportions . . . . .	473
D.8 Sample Size Estimates (Population Mean) . . . . .	475
D.9 Sample Size Estimates (Population Proportion) . . . . .	476
D.10 Sample Size Estimates (Two Population Means) . . . . .	476
D.11 Inference for Variances . . . . .	477
D.12 Goodness of Fit Tests . . . . .	479
D.13 Test for Independence in Contingency Tables . . . . .	479
D.14 ANOVA . . . . .	481
D.15 Wilcoxon Signed Rank Test for Paired Comparisons . . . . .	483
D.16 Wilcoxon Rank Sum Test for Independent Samples . . . . .	484

# Chapter 1

## Introduction

This volume provides the core material for an introductory course on statistical methodology. It contains the standard curriculum for an introductory course in statistics suitable for a undergraduate degree program in the quantitative sciences. A basic knowledge of calculus is assumed.

The volume is divided into three parts:

I Probability [Chapters 2-7]

II Statistics - Introduction [Chapters 8-11]

III Statistics - Inference [Chapters 12-29]

A separate volume *Introduction to Statistical Methodology (Exercises with Solution)* by the same author contains several hundred practice problems (with solutions) intended to give the reader practice in applying the statistical methods covered in this volume.

Statistics depends on probability theory, so this is covered first. Chapter 4 on “Random Variables” forms the core of this section. Part II constitutes an informal introduction to statistics, which emphasizes the management and summarization of data. Part III constitutes an introduction to formal inference, based on the confidence interval and the hypothesis test. The purpose of formal inference is to efficiently turn data into information while rigorously controlling for error. It is often said that when one looks for pattern one finds it. Inference addresses this problem.

In addition to the core statistics component, there is an emphasis on statistical computation. There is an extensive tutorial on the R statistical computing environment in Appendix C. In addition, throughout the volume, methodologies are demonstrated using R, sometimes in dedicated chapters. R is a GNU project, and may be downloaded free of charge from <http://www.r-project.org/>. The R website also hosts a considerable amount of supporting documentation. The main manual can be taken as *An Introduction to R*, by W. N. Venables, D. M. Smith and the R Core Team, and the student is advised to obtain this. There is also an extensive collection

of contributed manuals and tutorials at the same source. To access this material use the [manuals](#) link at the main page.

As mentioned, this material assumes a knowledge of calculus. Linear algebra is not a prerequisite, but occasionally a knowledge of matrix multiplication is assumed. However, it should be noted that linear algebra becomes essential at the intermediate level of statistical theory.

Appendix A contains extensive probability distribution tables. **R** can duplicate their function more conveniently, but they should still be understood, and their use should be considered part of the curriculum.

Appendix B contains a review of several mathematical topics which arise at various points in the volume. Section references are given as needed.

Appendix D contains a methodological summary, that can serve as a reference for the application of specific statistical procedures.

Although this volume is self contained, it may still be useful to have a second textbook which will offer more detail, as well as a source of sample problems and exercises. There is a large variety of such textbooks. They tend to be marketed towards specific disciplines. To some degree this determines their content, but in general the curriculum of this volume will be covered in almost all of them. The main exceptions to this will be "Simulation" (Chapter 6) and "Stochastic Processes" (Chapter 7), which are usually not covered in introductory statistics courses. Chapters 28 and 29 (ROC curves and simulation methods) tend to be covered after an introductory course.

The textbook *Probability and Statistics for Engineering and the Sciences* by Jay L. Devore, now in its 9th edition, can be highly recommended for statistical methodology. For probability and stochastic processes the textbook *Introduction to Probability Models* by Sheldon M. Ross, now in its 11th edition, can be highly recommended. It is not essential to have the latest edition of either textbook. Certainly, the large number of editions is an indication of their quality.

The textbook *Probability and Statistics for Computer Science* by James L. Johnson is considerably more advanced mathematically, and departs from the standard introductory statistics curriculum to a considerable degree. That being said, the material is of considerable interest, and could be considered as a reference for future study.

*Modern Applied Statistics with S* by W.N. Venables and B.D. Ripley, now in its 4th edition, can be highly recommended in general. The statistical methodology ranges from the intermediate to advanced, but contains much supplementary material that would be of interest to a reader of this volume, especially topics in advanced **R** programming. Note that the title makes reference to **S**, which is the proprietary statistical computing environment that preceded **R**. In fact, **R** has essentially the same functionality as **S** and most **S** code runs unaltered in **R**. This issue is discussed in the most recent edition.

We also note that *Springer* publishes a monograph series **Use R!** featur-

ing specialized applications of R, for example, on graphical models, bioinformatics, data mining, and so on.

The author should also acknowledge the adaption for this volume of some material from his own textbook *Approximate Iterative Algorithms* (A. Almudevar), published by *CRC Press*.

One of the notable qualities of statistics is its variety of character, in turn relying on mathematical theory, algorithm design, programming skill, and scientific problem-solving skills quite unrelated to any methodological canon. This volume was designed with that in mind.

Anthony Almudevar  
Rochester, NY  
2015

# Part I

# Probability

## Chapter 2

# Probability - Basic Theory

Probability is the mathematics of random occurrences. It plays a foundational role in statistics, and is, on its own, an important field in pure and applied mathematics.

We'll begin with an example.

**Example 2.1** Suppose a type of game involves choosing 3 of 5 balls at random, and that the balls are labeled 1,2,3,4 and 5. We wish to ask the following questions.

1. What is the probability of selecting both 1 and 5?
2. What is the probability that 4 is not selected?
3. Given that 5 is selected, what is the probability that 1 is also selected?

The first thing to do is to enumerate all possible outcomes:

123	124	125	134	135	145
234	235	245			
345					

There are 10 possibilities in all. We will use this list to answer the questions.

1. Note that items 125, 135, 145 have both 1 and 5. This is 3 out of 10 chances, so the correct probability is 3/10.
2. Note that items 123, 125, 135, 235, do not contain a 4. So the correct probability is 4/10.
3. Note that items 125, 135, 145, 235, 245, 345 contain a 5. Of these, 125, 135, 145 also contain a 1. So the probability that a 1 is selected *given that a 5 is selected* is 3 chances in 6, or 1/2. ■

## 2.1 Definitions

The following definition forms the mathematical basis for the characterization of a random occurrence.

**Definition 2.1** In order to calculate probabilities we begin with a *sample space*  $S$ . This is the set of all possible outcomes. A *random experiment* occurs when an element of  $S$  is selected at random. Suppose  $E$  is a subset of  $S$  (that is, all elements of  $E$  are also in  $S$ ). Then  $E$  is called an *event*, and  $P(E)$  is the *probability* that  $E$  occurs, or alternatively,  $P(E)$  is the probability that the outcome of the random experiment is in  $E$ . ■

**Example 2.2** In the previous example, we had sample space

$$S = \{123, 124, 125, 134, 135, 145, 234, 235, 245, 345\}.$$

We reasoned that event

$$\begin{aligned} E &= \{1 \text{ and } 5 \text{ are both selected}\} \\ &= \{125, 135, 145\} \end{aligned}$$

has probability

$$P(E) = 3/10. \quad \blacksquare$$

## 2.2 Set Algebra

Events are subsets of a sample space  $S$ . If  $A$  is a subset of  $S$  then we write

$$A \subset S.$$

The study of probability requires a proficiency in set algebra. A set algebra requires a *universe*, or *universal set*, of which all other sets under consideration are subsets. In probability, the sample space functions as the universe. Addition and multiplication are *binary* operators on real numbers, while *negation* is a *unary* operator. A similar system of operators is used in set algebra.

**Definition 2.2** Suppose  $A$  and  $B$  are two subsets of  $S$ . The *union* of  $A$  and  $B$  is written

$$A \cup B = \{\text{All elements of } S \text{ which are in } A \text{ OR } B\}.$$

The *intersection* of  $A$  and  $B$  is written

$$A \cap B = \{\text{All elements of } S \text{ which are in } A \text{ AND } B\}.$$

Notice that the difference in the definitions of union and intersection lies entirely in the words OR and AND. The *complement* of  $A$  is written

$$A^c = \{\text{All elements of } S \text{ which are NOT in } A\}. \quad \blacksquare$$

Note that the definition of a complement depends on the sample space.

**Example 2.3** Suppose

$$S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$

$$A = \{1, 2, 3, 4, 5\}$$

$$B = \{2, 4, 6, 8, 10\}.$$

Then, according to the above definitions

$$A \subset S$$

$$B \subset S$$

$$A \cup B = \{1, 2, 3, 4, 5, 6, 8, 10\}$$

$$A \cap B = \{2, 4\}$$

$$A^c = \{6, 7, 9, 10\}$$

$$B^c = \{1, 3, 5, 7, 9\}. \quad \blacksquare$$

Sometimes it will be convenient to define a set with no elements in it, roughly corresponding to the number zero in real number algebra.

**Definition 2.3** The *empty set*, denoted  $\emptyset$ , is the set containing no elements. It is considered a subset of the universe.  $\blacksquare$

The empty set may be taken as the complement of the sample space. Logically, we must have

$$S^c = \emptyset$$

and

$$S = \emptyset^c.$$

There is sometimes interest in determining the number of elements in a set. In set theory the somewhat more general notion of *cardinality* is usually used.

**Definition 2.4** The *cardinality* of a set  $A$ , denoted  $|A|$ , is the number of elements in the set. However, if the number of elements is infinite, cardinality may distinguish between different forms of infiniteness. In particular, a set  $A$  may be one of the following three types:

*Finite* if  $|A| < \infty$ .

*Countable* if  $|A| = \infty$ , but the elements can be listed as  $x_1, x_2, \dots$

*Uncountable* if  $|A| = \infty$  but the elements cannot be listed as  $x_1, x_2, \dots$

A set which is finite or countable is called *discrete*.  $\blacksquare$

The set of integers is countable, while the set of real numbers is uncountable. Note that some conventions refer to a finite set as countable. The term *discrete* removes this ambiguity.

**Definition 2.5** We say that two sets are *mutually exclusive* or *disjoint* if they have no elements in common. ■

Logically, we must have

$$A \cap B = \emptyset$$

if  $A$  and  $B$  are mutually exclusive.

**Example 2.4** To continue the previous example, if we have

$$\begin{aligned} C &= \{4, 5, 6\} \\ D &= \{1, 3\} \end{aligned}$$

then  $C$  and  $D$  are mutually exclusive, since they have no element in common ( $C \cap D = \emptyset$ ). ■

We can use the operations  $\cup$ ,  $\cap$  and complementation in more complicated expressions, as long as we keep track of the order in which the operations are done using parentheses.

**Example 2.5** Suppose

$$\begin{aligned} S &= \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\} \\ A &= \{1, 2, 3, 4, 5\} \\ B &= \{2, 4, 6, 8, 10\} \\ C &= \{4, 5, 7\}. \end{aligned}$$

Then

$$\begin{aligned} A \cup B \cup C &= \{1, 2, 3, 4, 5, 6, 7, 8, 10\} \\ A \cap B \cap C &= \{4\} \\ (A \cup B) \cap C &= \{4, 5\} \\ (A \cap B) \cup C &= \{2, 4, 5, 7\} \\ (A \cup C)^c &= \{6, 8, 9, 10\}. \end{aligned} \quad \blacksquare$$

**Definition 2.6** If a set consists of only one element, say  $E = \{5\}$  it is often referred to as a *singleton*. ■

**Definition 2.7** *De Morgan's Laws* state that

$$\begin{aligned} (A \cup B)^c &= A^c \cap B^c \\ (A \cap B)^c &= A^c \cup B^c. \end{aligned} \quad \blacksquare$$

## 2.3 Rules of Probabilities

Intuitively, we expect probability systems to obey the following rules, given sample space  $S$ .

1.  $P(S) = 1$
2.  $P(\emptyset) = 0$
3. If  $A$  and  $B$  are mutually exclusive, then

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

4. For any  $A$  and  $B$ ,

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

5.  $P(A^c) = 1 - P(A)$
6. If  $A \subset B$  then  $P(A) \leq P(B)$ .

Rules 1 and 2 assert that the outcome of the random experiment consists of one and only one element of  $S$ .

Rule 3 asserts that probabilities are additive. The probability that the Boston Red Sox or the New York Yankees win the World Series is equal to the probability that the Boston Red Sox win the world series plus the probability that the New York Yankees win the world series.

Rule 4 is similar to Rule 3, except it takes into account the possibility that both  $A$  and  $B$  may occur if they are not mutually exclusive. The formula essentially corrects for double counting the elements that are in both  $A$  and  $B$ .

Rule 5 gives an intuitive rule for the probability of a complement.

Rule 6 conforms to our expectation that if event  $A$  is contained in  $B$  then the probability of  $A$  cannot be larger than the probability of  $B$ .

**Example 2.6** Suppose

$$\begin{aligned} S &= \{1, 2, 3, 4, 5\} \\ A &= \{1, 3\} \\ B &= \{3, 4, 5\}. \end{aligned}$$

Suppose we wish to calculate

$$\begin{aligned} P(A \cup B) &= P(\{1, 3\} \cup \{3, 4, 5\}) \\ &= P(\{1, 3, 4, 5\}). \end{aligned}$$

By Rule 3 we should have

$$\begin{aligned} P(A) &= P(\{1, 3\}) \\ &= P(\{1\}) + P(\{3\}) \\ P(B) &= P(\{3, 4, 5\}) \\ &= P(\{3\}) + P(\{4\}) + P(\{5\}) \end{aligned}$$

and

$$\begin{aligned} P(A \cap B) &= P(\{1, 3\} \cap \{3, 4, 5\}) \\ &= P(\{3\}). \end{aligned}$$

This gives, by Rule 4,

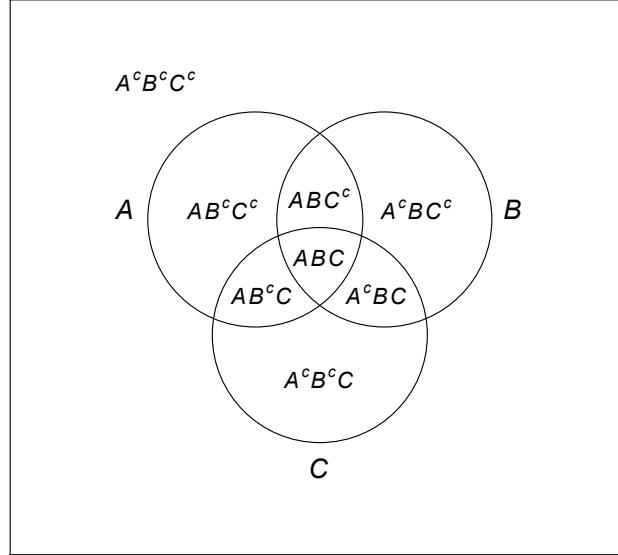
$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= (P(\{1\}) + P(\{3\})) + (P(\{3\}) + P(\{4\}) + P(\{5\})) - (P(\{3\})) \\ &= P(\{1\}) + P(\{3\}) + P(\{4\}) + P(\{5\}) \\ &= P(\{1, 3, 4, 5\}) \end{aligned}$$

which gives the correct result. Note that the main effect of the quantity  $-P(A \cap B)$  in Rule 4 in the above calculation is to prevent  $P(\{3\})$  from being counted twice. ■

## 2.4 Venn Diagram

The *Venn diagram* is a schematic device used to represent interactions between multiple sets. Suppose we are given two events  $A, B$ . In one sense these events generate 4 new events,  $AB, AB^c, A^cB$  and  $A^cB^c$ . Similarly, 3 events generate 8 new events, and in general  $n$  events generate  $2^n$  events. The Venn diagram represents events as sets, which overlap so as to represent all possible intersections of the sets and their complements. Figure 2.1 is a typical example for 3 sets  $A, B$  and  $C$ . The total region is partitioned into the 8 regions induced by the intersections. The regions may be annotated by probabilities or counts.

A Venn diagram may be useful for Example 2.9 below.

Figure 2.1: Venn diagram for 3 sets  $A, B, C$ .

## 2.5 Independence

If we toss a coin twice, we don't expect that the outcome of the first toss will tell us anything about the outcomes of the second toss. When two events have no such probabilistic relationship we say they are *independent*. Mathematically, if two events  $A$  and  $B$  are independent then

$$P(A \cap B) = P(A)P(B). \quad (2.1)$$

Sometimes the notation  $A \perp B$  is used to denote the independence relationship (2.1).

**Example 2.7** Suppose we toss a coin twice. Assume that the tosses are independent. Then if we set

$$\begin{aligned} A &= \{\text{first toss is a head}\} \\ B &= \{\text{second toss is a head}\} \end{aligned}$$

then

$$A \cap B = \{\text{both tosses are heads}\}.$$

Intuitively we have

$$\begin{aligned} P(A) &= 1/2 \\ P(B) &= 1/2. \end{aligned}$$

We can therefore write

$$\begin{aligned} P(\{\text{both tosses are heads}\}) &= P(A \cap B) \\ &= P(A)P(B) \\ &= 1/2 \times 1/2 \\ &= 1/4. \end{aligned}$$

More generally we have

$$\begin{aligned} P(\{\text{3 coin tosses are heads}\}) &= 1/2 \times 1/2 \times 1/2 \\ &= 1/8 \end{aligned}$$

and eventually

$$\begin{aligned} P(\{\text{N coin tosses are heads}\}) &= 1/2 \times \dots \times 1/2 \\ &= (1/2)^N. \blacksquare \end{aligned}$$

**Example 2.8** A roulette wheel has 38 slots, of which 18 are red, 18 are black and 2 are green. After a spin, a ball lands at random in one of the slots. Assume all spins are independent. What is the probability that the ball lands in red in at least one of the next two spins?

Let

$$\begin{aligned} A &= \{\text{at least one red}\} \\ B_1 &= \{\text{first spin is not red}\} \\ B_2 &= \{\text{second spin is not red}\}. \end{aligned}$$

Logically, we must have

$$A^c = (B_1 \cap B_2).$$

We also have

$$\begin{aligned} P(B_1) &= 20/38 \\ P(B_2) &= 20/38 \end{aligned}$$

which gives

$$\begin{aligned} P(A) &= 1 - P(A^c) \\ &= 1 - P(B_1 \cap B_2) \\ &= 1 - P(B_1)P(B_2) \\ &= 1 - (20/38)^2 \\ &= 0.723 \end{aligned}$$

since  $B_1$  and  $B_2$  are independent.  $\blacksquare$

The notion of independence can be extended to more than two sets, but requires some care.

**Definition 2.8** Suppose we are given events  $A_1, A_2, \dots, A_n$ . We say these sets are *independent* if

$$P(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_m}) = P(A_{i_1})P(A_{i_2}) \dots P(A_{i_m})$$

for every subcollection  $A_{i_1} \cap A_{i_2}, \dots, \cap A_{i_m}$  of the original sets. ■

For example, three sets  $A_1, A_2, A_3$  are independent if and only if all the following hold:

$$\begin{aligned} P(A_1 \cap A_2) &= P(A_1)P(A_2) \\ P(A_2 \cap A_3) &= P(A_2)P(A_3) \\ P(A_1 \cap A_3) &= P(A_1)P(A_3) \\ P(A_1 \cap A_2 \cap A_3) &= P(A_1)P(A_2)P(A_3). \end{aligned}$$

If we may only claim that  $P(A_i \cap A_j) = P(A_i)P(A_j)$  for every pair of events, we say that the events are *pairwise independent*.

It is necessary to insist on the product rule for all subcollections of sets. This is demonstrated in the next example.

**Example 2.9** Consider three sets  $A, B, C$ . We may construct all joint probabilities by specifying suitable probability values for all eight regions of the Venn diagram. Suppose we set  $P(ABC) = 1/64$ ,  $P(AB^cC^c) = P(A^cBC^c) = P(A^cB^cC) = 15/64$  and  $P(A^cB^cC^c) = 18/64$ . This gives  $P(A) = P(B) = P(C) = 1/4$ , and so  $P(ABC) = P(A)P(B)P(C)$ . On the other hand,  $P(AB) = 1/64 \neq P(A)P(B) = 1/16$ , and so  $A$  and  $B$  are not independent.

Conversely, pairwise independence does not imply independence. For example, if  $A \perp B$ ,  $P(A) = P(B) = 1/2$  and  $C = AB \cup A^cB^c$ , then it is easily verified that we also have  $A \perp C$  and  $B \perp C$ , but that  $P(ABC) \neq P(A)P(B)P(C)$ . ■

## 2.6 Conditional Probability

If  $A$  and  $B$  are two events, then  $P(A)$  is the probability that  $A$  occurs. If, however, we somehow know that  $B$  has occurred, this may alter the probability that  $A$  occurs. We can estimate, for example, the probability that a given individual develops cancer during the next 10 years (event  $A$ ). If, however, we know that a close relative has developed cancer (event  $B$ ), this alters the probability.

**Definition 2.9** The *conditional* probability of  $A$  given  $B$  is

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

**Example 2.10** In a previous example, we considered a random experiment in which 3 balls were randomly selected from 5 labeled 1,2,3,4,5. All possible outcomes are given below

123	124	125	134	135	145
234	235	245			
345					

It was deduced that the probability that a 1 is selected *given* that a 5 is selected is  $1/2$ , since there are 6 outcomes in which a 5 is selected, and of these, 3 outcomes have a 1. If we use the definition of conditional probability we have

$$\begin{aligned}
 A &= \{1 \text{ is selected}\} \\
 B &= \{5 \text{ is selected}\} \\
 P(1 \text{ is selected } | 5 \text{ is selected}) &= \frac{P(A \cap B)}{P(B)} \\
 &= \frac{P(1 \text{ and } 5 \text{ are selected })}{P(5 \text{ is selected })} \\
 &= \frac{P(125,135,145)}{P(125,135,145,235,245,345)} \\
 &= \frac{3/10}{6/10} \\
 &= 1/2,
 \end{aligned}$$

which gives the same answer. ■

**Example 2.11** The following table is an example of a *contingency table*.

Marital status	Caffeine Consumption (mg/day)				Total
	0	1-150	151-300	> 300	
Married	652	1537	598	242	3029
Divorced, separated or widowed	36	46	38	21	141
Single	218	327	106	67	718
Total	906	1910	742	330	3888

We are often interested in calculating conditional probabilities from contingency tables. For example, if we want to know the probability that a single subject consumes 0 caffeine given that he/she is single, we set

$$\begin{aligned}
 A &= \{\text{consumes 0 caffeine}\} \\
 B &= \{\text{subject is single}\}
 \end{aligned}$$

then

$$\begin{aligned}
 P(A | B) &= \frac{P(\text{subject consumes 0 caffeine and is single})}{P(\text{subject is single})} \\
 &= \frac{218/3888}{718/3888} \\
 &= 218/718 \\
 &= 0.30. \quad \blacksquare
 \end{aligned}$$

The notion of independence is related to conditional probabilities. Intuitively, if  $A \perp B$ , then we would not expect the occurrence of  $B$  to affect the probability of  $A$ , that is, we expect  $P(A | B) = P(A)$ , and similarly  $P(B | A) = P(B)$ . That this is the case is verified in the next theorem

**Theorem 2.1** Given events  $A, B$ ,

$$P(A | B) = P(A) \text{ and } P(B | A) = P(B)$$

if and only if  $A \perp B$ . ■

*Proof.* First, suppose  $A$  and  $B$  are independent, in which case we have

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

from which it follows that

$$\begin{aligned}
 P(A | B) &= \frac{P(A \cap B)}{P(B)} \\
 &= \frac{P(A)P(B)}{P(B)} \\
 &= P(A),
 \end{aligned}$$

and by a similar argument  $P(B | A) = P(B)$ .

Next, suppose  $P(A | B) = P(A)$ . Then

$$\begin{aligned}
 P(A) &= P(A | B) \\
 &= \frac{P(A \cap B)}{P(B)}.
 \end{aligned}$$

This may be rearranged to give  $P(A \cap B) = P(A)P(B)$ , or equivalently,  $A \perp B$ . The same follows by assuming  $P(B | A) = P(B)$ . □

## 2.7 The Law of Total Probability

Suppose we have a collection of mutually exclusive sets  $A_1, \dots, A_n$  and we let

$$A = \bigcup_{i=1}^n A_i$$

We say that the collection of sets *partitions*  $A$ . It is often useful to calculating the probability of an event  $E$  by first conditioning on each element of a partition.

**Definition 2.10** Suppose  $A_1, \dots, A_n$  is a partition of sample space  $S$ . The *law of total probability* states that for any other event  $E$

$$P(E) = P(E | A_1)P(A_1) + P(E | A_2)P(A_2) + \dots + P(E | A_n)P(A_n).$$

For a conditional probabilities we have

$$\begin{aligned} P(E | B) &= P(E | A_1B)P(A_1 | B) + P(E | A_2B)P(A_2 | B) + \\ &\dots + P(E | A_nB)P(A_n | B). \end{aligned} \quad (2.2)$$

provided  $P(B) > 0$ . ■

**Example 2.12** Suppose at a given locus a genotype has two alleles  $A$  and  $a$ , with population frequencies  $p_A + p_a = 1$ . If the genotype of the mother is  $Aa$  what is the probability that the offspring has genotype  $aa$ ?

We know the mother's genotype but not the father's genotype. Let

$$E = \{ \text{The offspring has genotype } aa \}.$$

We then *condition* on the father's genotype, that is, we calculate the probability of  $E$  conditional on the father having a specific genotype, for each of the three possible genotypes. This gives, under Mendel's laws of inheritance

$$\begin{aligned} P(E | aa) &= 1/2 \\ P(E | Aa) &= 1/4 \\ P(E | AA) &= 0. \end{aligned}$$

To calculate the probability of  $E$  we then need the probability that the father is a given genotype for each of the three genotypes. If we assume Hardy-Weinberg equilibrium this will be

$$\begin{aligned} P(aa) &= p_a^2 \\ P(Aa) &= 2p_Ap_a \\ P(AA) &= p_A^2. \end{aligned}$$

Since the father's genotype can be one and only one of the three genotypes  $aa$ ,  $Aa$  and  $AA$  we can use the law of total probability to get

$$\begin{aligned} P(E) &= P(E | aa)P(aa) + P(E | Aa)P(Aa) + P(E | AA)P(AA) \\ &= 1/2 \times p_a^2 + 1/4 \times 2p_Ap_a + 0 \times p_A^2 \\ &= \frac{p_a^2 + p_Ap_a}{2} \\ &= \frac{p_a(p_a + p_A)}{2} \\ &= \frac{p_a}{2}. \end{aligned} \quad ■$$

## Chapter 3

# Combinatorics and Equiprobable Distributions

Many probability models are constructed as “random selections” from some collection of objects. What is usually implied is formally an equiprobable sample space, and many real-life random processes are designed to be precisely this, especially games of chance.

When this is the case, probability computations are largely a process of enumeration, and typically involve two numbers. Given a sample space  $S$ , under the equiprobable model the probability of an event  $E \subset S$  is

$$P(E) = \frac{N}{D} \quad (3.1)$$

where  $D$  is the total number of outcomes in  $S$ , and  $N$  is the number of outcomes in  $E$ .

**Example 3.1** A committee of 2 students is to be randomly selected from 4 eligible students. What is the probability that any given student is on the committee?

Label the students  $A, B, C, D$  and we’ll evaluate the probability that  $A$  is on the committee. We can enumerate all pairs:

$$S = \{AB, AC, AD, BC, BD, CD\}$$

and the set of interest:

$$E = \{AB, AC, AD\}.$$

Then using (3.1) we have  $N = 3$ ,  $D = 6$  and therefore

$$P(E) = \frac{N}{D} = \frac{3}{6} = 1/2. \quad \blacksquare$$

### 3.1 Counting Rules

*Combinatorics* is the branch of mathematics which studies the structure of finite or countably infinite discrete structures. *Enumerative combinatorics*

concerns the problem of enumerating (counting) the number of objects of a certain type. It turns out that a very large number of enumeration problems can be solved using a relatively small number of core principles.

To illustrate, in Example 3.1 we counted  $D = 6$  possible committees, and  $N = 3$  committees which include student  $A$ . Note that no distinction is made between the 2 committee members. Suppose instead that one committee member is designated president. If we adopt the notational convention that the president is listed first, then we can see that the selections  $AB$  and  $BA$  are now different, so the set of all outcomes is now

$$S = \{AB, BA, AC, CA, AD, DA, BC, CB, BD, DB, CD, DC\},$$

so that the quantity  $D$  in (3.1) is now 12. The probability that  $A$  is on the committee is again calculated by enumerating the relevant outcomes

$$E = \{AB, BA, AC, CA, AD, DA\}$$

giving  $N = 6$  so that  $P(E) = 6/12 = 1/2$ , which is the same probability as in Example 3.1. However, if we ask instead the probability that  $A$  is president, we have a new set

$$F = \{AB, AC, AD\},$$

which is not the same set as  $E$  in Example 3.1, despite the apparent equivalence. We now have  $N = 3$ , so that the probability that  $A$  is president is  $P(F) = 3/12 = 1/4$ .

The important distinction here is between *ordered* and *unordered* selections:

**Definition 3.1** Suppose we are given a finite set  $S$ . An *ordered selection* of size  $n$  from  $S$  is a selection of  $n$  distinct objects from  $S$  paired with the order of selection. An *unordered selection* of size  $n$  from  $S$  is a selection of  $n$  distinct objects from  $S$  without reference to any order of selection. ■

Intuitively, the order of an ordered selection need not represent any sequential process, and may represent any ranking imposed on the objects. In the previous example, designating one committee member as president has the effect of imposing an ordering. Thus, a committee of 2 students is an unordered selection, while a committee defined by designated positions is an ordered selection.

### 3.1.1 Rule of Product

The most general counting rule is the *rule of product*.

**Definition 3.2** Suppose a procedure can be broken down into  $K$  tasks, and there are always  $n_i$  distinct ways to perform the  $i$ th task, for  $i = 1, \dots, K$ . Then there are  $n_1 \times n_2 \times \dots \times n_K$  distinct ways to perform the entire procedure. ■

**Example 3.2** How many "words" have three letters, with exactly one vowel, located in the center, such that no letter is used twice? We can think of a word as a procedure with three tasks, so  $K = 3$ . The first task is to select the first letter, which must be a consonant, of which there are 21. This means  $n_1 = 21$ . The second task is to select the second letter, a vowel, of which there are five, so that  $n_2 = 5$ . Note that a vowel must always be different from a consonant, so we may select any vowel. The third task is to select the final consonant. Since all letters must be different, we cannot select the first letter over again. This leaves 20 consonants to choose from, so that  $n_3 = 20$ . The number of words matching our conditions is therefore

$$n_1 \times n_2 \times n_3 = 21 \times 5 \times 20 = 2100. \quad \blacksquare$$

### 3.1.2 Enumeration of Ordered and Unordered Selections

An important quantity in counting rules is the *factorial*, which we define next

**Definition 3.3** Let  $n$  be any nonnegative integer. The *factorial* of  $n$ , written  $n!$ , ( $n$ -factorial) is defined as follows. For  $n \geq 1$

$$n! = n \times (n - 1) \times \dots \times 2 \times 1$$

and for  $n = 0$  we set

$$0! = 1. \quad \blacksquare$$

The factorial is important in counting problems, since it gives the number of ways to arrange in order  $n$  distinct objects, since, by the rule of product, there are  $n$  ways to select the first item,  $n - 1$  ways to select the second items, all the way down to 1 way to select the final item. An ordering of  $n$  objects is called a *permutation*.

The same logic can be used to enumerate ordered selections of size  $k$ . There are  $n$  ways to select the first object,  $n - 1$  ways to select the second objects, and  $n - k + 1$  ways to select the final number. We denote this quantity

$$P(n, k) = n \times (n - 1) \times \dots \times (n - k + 1).$$

This value can be made somewhat more convenient using factorials in the following way

$$\begin{aligned} P(n, k) &= n \times (n - 1) \times \dots \times (n - k + 1) \\ &= n \times (n - 1) \times \dots \times (n - k + 1) \times \frac{(n - k)!}{(n - k)!} \\ &= \frac{n \times (n - 1) \times \dots \times (n - k + 1) \times (n - k)!}{(n - k)!} \\ &= \frac{n!}{(n - k)!}. \end{aligned}$$

Note that the term *permutation* is generally applied to any ordered selection. A permutation of  $n$  objects may be thought of as an ordered selection of  $k = n$  from  $n$  objects, accordingly,

$$P(n, n) = \frac{n!}{(n - n)!} = n!.$$

Next, we consider the enumeration of an unordered selection of  $k$  from  $n$  objects, referred to as a *combination*. Denote the solution  $C(n, k)$ . Suppose we know this number, and we want to deduce  $P(n, k)$ . Note that an ordered selection of  $k$  from  $n$  is equivalent to an unordered selection of  $k$  from  $n$ , paired with a permutation of the  $k$  selection objects, of which there are  $k!$ . Therefore, by the rule of product we have

$$P(n, k) = C(n, k)k!,$$

or equivalently,

$$C(n, k) = \frac{P(n, k)}{k!} = \frac{n!}{(n - k)!k!} = \binom{n}{k},$$

which is generally known as the *binomial coefficient*, as well as  $n$  choose  $k$ .

We summarize the quantities  $P(n, k)$  and  $C(n, k)$  in the following definition.

**Definition 3.4** An ordered section of  $k$  from  $n$  distinct objects is a *permutation*. The number of permutations of  $k$  from  $n$  is

$$P(n, k) = \frac{n!}{(n - k)!}.$$

An unordered section of  $k$  from  $n$  distinct objects is a *combination*. The number of combinations of  $k$  from  $n$  is

$$C(n, k) = \binom{n}{k} = \frac{n!}{(n - k)!k!}.$$

This quantity is generally known as the *binomial coefficient*, as well as  $n$  choose  $k$ . ■

**Example 3.3** To count the number of permutations of the letters  $abc$ , we may calculate

$$\begin{aligned} 3! &= 3 \times 2 \times 1 \\ &= 6. \end{aligned}$$

So there are 6 permutations. We can list them explicitly:

$$abc \ acb \ bac \ bca \ cab \ cba. \quad \blacksquare$$

**Example 3.4** In poker, a hand consists of 5 cards chosen from a standard deck of 52. We can use the binomial coefficient to count the number of possible hands by setting  $n = 52$ ,  $k = 5$  to get

$$\begin{aligned}\binom{52}{5} &= \frac{52 \times 51 \times 50 \times 49 \times 48}{5 \times 4 \times 3 \times 2 \times 1} \\ &= 2,598,960\end{aligned}$$

as the number of possible hands. ■

**Example 3.5** Suppose we want to calculate the probability of being dealt “4 of a kind” in poker. A hand is “4 of a kind” if it has 4 of one single rank. The sample space is  $S$ , the set of all selections of 5 cards from 52. From the Example 3.4 we have denominator

$$D = 2,598,960.$$

If we let  $E$  be the set of all “4 of a kind”, we can use the rule of product to calculate  $N$ , the size of  $E$ . We can construct a “4 of a kind” using  $K = 2$  tasks. We first select the 4 cards of identical rank. Since there are 13 ranks we may set  $n_1 = 13$ . After the first task is complete we need to select the remaining single card. We can’t select any of the 4 cards used for task 1, which leaves  $52-4=48$  remaining cards for the second task, which means  $n_2 = 48$ , giving numerator

$$\begin{aligned}N &= n_1 \times n_2 \\ &= 13 \times 48 \\ &= 624.\end{aligned}$$

Finally, we have

$$\begin{aligned}P(E) &= \frac{N}{D} \\ &= \frac{624}{2,598,960} \\ &= \frac{1}{4165}\end{aligned}$$

so that there is a one in 4,165 chance of being dealt a “4 of a kind”. Of course, we have assumed that all hands are equally likely to be dealt. ■

### 3.1.3 Permutations with Repetitions - The Multinomial Coefficient

Suppose we wish to enumeration the permutations of  $n$  objects, which need not be entirely distinct. For example, suppose we wish to order the labels  $A, A, B, C$ . Since the first two labels are identical, the number of permutations will not be obtained form the quantity  $P(4, 4)$ . For this type of

enumeration problem, we may employ a device frequently used in combinatorics, in particular, the temporary labeling of identical objects. In this case, we label the two  $A$  elements, giving  $A_1, A_2, B, C$ , of which there are  $P(4, 4) = 4! = 24$ ,

$$A_1A_2BC, A_2A_1BC, A_1BA_2C, A_2BA_1C, \dots, CBA_1A_2, CBA_2A_1. \quad (3.2)$$

If we enumerate the permutations of  $A, A, B, C$  we would obtain a list such as,

$$AABC, ABAC, \dots, CBAA. \quad (3.3)$$

Clearly, there must be a natural mapping between these two lists:

$$\begin{aligned} A_1A_2BC, A_2A_1BC, &\iff AABC \\ A_1BA_2C, A_2BA_1C, &\iff ABAC \\ &\vdots \\ CBA_1A_2, CBA_2A_1, &\iff CBAA. \end{aligned}$$

We should then be able to construct enumeration (3.2) from (3.3) in the following way: for each object in (3.3) (say  $ABAC$ ) generate objects in (3.2) by assigning temporary labels to the identical elements in all possible ways (that is,  $A_1BA_2C$  and  $A_2BA_1C$ ). The number of ways to assign temporary labels in this way is equal to the number of permutations of  $k = 2$  identical elements, in this case  $k! = 2$ . Therefore, the number of permutations of  $A_1, A_2, B, C$  must be  $k! = 2$  times the number of permutations of  $A, A, B, C$ . But we know the number of permutations of 4 distinct objects (such as  $A_1, A_2, B, C$ ) is  $4! = 24$ . This means the number of permutations of  $A, A, B, C$  must be  $24/2 = 12$ .

At this point we could state a general principle: if we are given  $n$  elements, of which all are distinct except for  $k$  identical elements, then the number of permutations of these elements is

$$N_{perm} = \frac{n!}{k!}.$$

However, a greater degree of generality does not require much extra work. Suppose we are given  $n$  elements of  $m$  distinct types. Suppose there are  $n_i$  elements of type  $i$ ,  $i = 1, \dots, m$ , so that

$$n_1 + \dots + n_m = n.$$

We may similarly apply temporary labels independently to each type. For example, if we are given  $n = 6$  elements  $AABCCC$  we have  $m = 3$  distinct element types, with  $n_1 = 2, n_2 = 1, n_3 = 3$ , with  $n_1 + n_2 + n_3 = 2 + 1 + 3 = 6 = n$ . We then apply temporary labels  $A_1A_2B_1C_1C_2C_3$ . There are  $6! = 720$  permutations of the temporarily labeled elements. As in the previous

example, we may successively divide by the number of permutations of the temporary labels within each element type, giving

$$N_{perm} = \frac{n!}{n_1! \times n_2! \times n_3!} = \frac{6!}{2! \times 1! \times 3!} = \frac{720}{2 \times 1 \times 6} = 60.$$

At this point we have the following general principle.

**Definition 3.5** The number of permutations of  $n$  elements of  $m$  distinct types, with  $n_i$  elements of type  $i$ ,  $i = 1, \dots, m$ , is given by the *multinomial coefficient*:

$$\binom{n}{n_1, \dots, n_m} = \frac{n!}{\prod_{i=1}^m n_i!} \quad (3.4)$$

Note that the multinomial coefficient reduces to the binomial coefficient for  $m = 2$ . ■

**Example 3.6** How many permutations of the elements *SSSFFFF* are there?

There are two ways to approach this problem. Note that a permutation of *SSSFFFF* is equivalent to a selection of 3 of 7 positions for the elements *S*. That is, for permutation *SFFSFSF* the *S* elements occupy positions 1,4 and 6 of the available 7 positions. So,

$$N_{perm} = C(7, 3) = \binom{7}{3} = \frac{7!}{3! \times 4!} = 35.$$

On the other hand, we have  $n = 7$  elements of  $m = 2$  types, with  $n_1 = 3$  and  $n_4$ . Therefore, using multinomial (or binomial) coefficient (3.4) we have

$$N_{perm} = \binom{n}{n_1, n_2} = \frac{7!}{3! \times 4!} = 35.$$

We (necessarily) get the same answer using each method. Enumeration problems often have this character. ■

The following example illustrates the manner in which the various enumeration principles are combined within a single enumeration method.

**Example 3.7** An urn contains 50 red balls (numbered 1 to 50) and 50 green balls (number 1 to 50). 4 balls are chosen at random. The total number of ways to select 4 balls from 100 is

$$\begin{aligned} \binom{100}{4} &= (100 \times 99 \times 98 \times 97)/(4 \times 3 \times 2 \times 1) \\ &= (100/4) \times (99/3) \times (98/2) \times (97/1) \\ &= 25 \times 33 \times 49 \times 97 \\ &= 3,921,225. \end{aligned}$$

We will then find the following probabilities, making use of denominator  $D = 3,921,225$ .

**Problem (A)** *There are two pairs of balls which have the same number.* Selecting 4 balls with two pairs with the same number is equivalent to selecting 2 numbers from 50. There are

$$\binom{50}{2} = (50 \times 49)/(2 \times 1) = 1225$$

ways to do this. Therefore

$$P(\text{two pairs with the same number}) = 1225/3921225 \approx 0.00031.$$

**Problem (B)** *The balls are of the same color and consecutively numbered.* Use the rule of product. We make the selection in two stages. First, select the color, then select the sequence. In detail,

1. Select color.  $N_1 = 2$  (there are 2 colors)
2. Select sequence.  $N_2 = 47$  (a consecutive sequence of 4 numbers, selected from numbers 1 to 50, can start with any number from 1 to 47)

There are  $N_1 \times N_2 = 2 \times 47 = 94$  ways to select 4 balls such that all are of the same color and the numbers are consecutive.

$$P(\text{same color and the numbers are consecutive}) = 94/3921225 \approx 2.4 \times 10^{-5}.$$

**Problem (C)** *Two (and only two) balls have the same number.* Use the rule of product. First select the pair with the same number. Then select 2 different numbers from those remaining. Finally, color the last two balls.

1. Select number for matching pair.  $N_1 = 50$  (there are 50 numbers which can be paired)
2. Select two remaining numbers.  $N_2 = 1176$  (there are 49 choose 2 ways to select two numbers from 49, which equals  $49 \times 48/2 = 1176$ )
3. Select the color combinations for the remaining two balls.  $N_3 = 4$  (the combinations can be red/green, green/red, red/red, green/green)

There are  $N_1 \times N_2 \times N_3 = 50 \times 1176 \times 4 = 235,200$  ways to select 4 balls such that two, and only two, have the same number.

$$P(\text{two and only two have the same number}) = 235200/3921225 \approx 0.05998.$$

We offer an alternative solution to this problem, based on the complement rule:

$$\begin{aligned} P(\text{at least two balls with the same number}) \\ = 1 - P(\text{no two balls have the same number}). \end{aligned}$$

Use the rule of product to calculate the number of selections for which no two balls have the same number. First, select 4 distinct numbers from 50, then color the numbers.

1. Select four distinct numbers.  $N_1 = 230,300$  (there are '50 choose 4' = 230,300 ways to choose 4 distinct numbers from 50)
2. Select color combination for 4 distinct numbers.  $N_2 = 16$  (each of 4 numbers can be red or green, so there are  $2 \times 2 \times 2 \times 2 = 16$  color combinations.)

There are  $N_1 \times N_2 = 230,300 \times 16 = 3,684,800$  ways to select 4 balls such no two have the same number.

$$P(\text{no two balls have the same number}) = 3684800/3921225 \approx 0.9397063.$$

So, let

$$\begin{aligned} A &= \{\text{two pairs with the same number}\} \\ B &= \{\text{two and only two have the same number}\} \\ C &= \{\text{at least two balls with the same number}\} \end{aligned}$$

We solve the problem by evaluating

$$P(B) = P(C) - P(A) \approx (1 - 0.9397063) - 0.00031 \approx 0.05998.$$

Note that we get  $P(A)$  from **Problem (A)** of this example. ■

## Chapter 4

# Random Variables

We begin with the basic definition:

**Definition 4.1** A *random variable* (RV) is a numerical outcome  $X \in \mathbb{R}$  associated with a random experiment. Associated sets of the form

$$A = \{X \in E\}$$

for subsets  $E \subset \mathbb{R}$  are assumed to conform to the rules of probability given in Section 2.3. The random variable may also be considered a random experiment with a sample space of outcomes  $S_X \subset \mathbb{R}$ , in which case  $X$  must be a value in  $S_X$ .

The *distribution* of a RV  $X$  is expressible as probabilities of the form  $P(X \in E)$ . We may adopt the shorthand

$$P_X(E) = P(X \in E).$$

■

If  $S_X$  consists of integers, we use the shorthand

$$p_i = P(X = i).$$

In this case the *distribution* of  $X$  consists of the values of  $p_i$  for all  $i \in S_X$ .

**Example 4.1** Suppose we toss a coin three times. Assume all possible outcomes have equal probability. The outcomes are

H H H   H H T   H T H   H T T   T H H   T H T   T T H   T T T.

Let  $X$  be the number of heads. Then  $S_X = \{0, 1, 2, 3\}$ . The distribution of  $X$  is then

$$\begin{aligned} p_0 &= P(X = 0) &= P(T T T) &= 1/8 \\ p_1 &= P(X = 1) &= P(T T H, T H T, H T T) &= 3/8 \\ p_2 &= P(X = 2) &= P(H H T, H T H, T H H) &= 3/8 \\ p_3 &= P(X = 3) &= P(H H H) &= 1/8. \end{aligned}$$

Note that

$$p_0 + p_1 + p_2 + p_3 = 1$$

as expected.

■

## 4.1 Types of Sample Spaces

The form of sample space  $S_X$  has important implications for probabilistic analysis. Recalling Definition 2.4 a sample space may be finite, countable or uncountable.

Some examples of countable sample spaces are sets of integers, such as

$$\{1, 2, \dots\} \text{ or } \{0, 1, 2, \dots\} \text{ or } \{\dots, -2, -1, 0, 1, 2, \dots\}.$$

The best known examples of uncountable sample spaces are the set of real numbers, or intervals of real numbers such as  $[0, 1]$ .

It serves our purpose to categorize random variables as *discrete* if  $S_X$  is finite or countable, and *continuous* if  $S_X$  is uncountable.

There are some random variables which are neither entirely discrete or continuous, for example, the waiting time in a bank queue, if there is a probability greater than 0 that the waiting time will be 0.

## 4.2 Probability Mass Functions and Densities

A random variable is defined by its sample space and either a *probability mass function* (PMF) for discrete random variables or a *density function* for continuous random variables. This approach guarantees that a RV satisfies Definition 4.1 in full, in particular, all events involving the RV will conform to the probability rules of Section 2.3. Some authors refer to a PMF also as a density, and more advanced probability theory tends to unify definitions for discrete and continuous random variables (remember that a random variable need not be either type). However, calculation methods are different for each type, so there is some advantage to considering the two cases separately.

We have the following definitions:

**Definition 4.2** For a discrete random variable  $X$  with sample space  $S_X$  a *probability mass function* (PMF) assigns a number  $p_X(x) \in [0, 1]$  to each element  $x \in S_X$  so that

$$\sum_{x \in S_X} p_X(x) = 1$$

A *density function* is a function  $f(x)$  with the following properties.

1. The function  $f(x)$  is always greater than or equal to zero.
2. The total area under the function is one.

A random variable  $X$  is continuous if and only if there is a density function  $f_X$  with the property that

$$P(a \leq X \leq b) = \int_a^b f_X(x) dx = \text{Area under density function between } a \text{ and } b.$$

The *support*  $\mathcal{S}_X$  of a discrete random variable consists of all numbers  $x \in \mathbb{R}$  for which  $p_X(x) > 0$ , and the *support*  $\mathcal{S}_X$  of a continuous random variable consists of all numbers  $x \in \mathbb{R}$  for which  $f_X(x) > 0$ . ■

The support is necessarily a subset of the sample space. It is usually not necessary to distinguish between the sample space and support of a random variable, but the difference may become important when several random variables are defined on the same sample space.

**Example 4.2** Suppose we toss a coin 3 times, letting  $X$  equal the number of heads. We consider all sequences of 3  $H$  or  $T$  labels to be an outcome, so that the sample space consists of the 8 possible sequences

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}.$$

If we assume that each outcome is equally likely (equivalently, has probability  $1/8$ ), then we can calculate the following probabilities:

$$\begin{aligned} P(X = 0) &= P(\{TTT\}) = 1/8 \\ P(X = 1) &= P(\{HHT, HTH, THH\}) = 3/8 \\ P(X = 2) &= P(\{HTT, THT, TTH\}) = 3/8 \\ P(X = 3) &= P(\{HHH\}) = 1/8. \end{aligned}$$

This defines the PMF  $p_X$ :

$$\begin{aligned} p_X(0) &= 1/8 \\ p_X(1) &= 3/8 \\ p_X(2) &= 3/8 \\ p_X(3) &= 1/8. \end{aligned}$$

■

#### 4.2.1 Normalization of PMFs and Densities

We are sometimes given positive discrete or continuous functions which are intended to represent PMFs or densities, but which do not sum or integrate to one, that is  $P(S_X) \neq 1$ . The term *normalization* refers to the scalar adjustment of such a function so that  $P(S_X) = 1$ . For the discrete case, if  $g(x) > 0$  for  $x \in S_X$  then

$$p(x) = \frac{g(x)}{\sum_{x' \in S_X} g(x')} \quad (4.1)$$

is a PMF, and for the continuous case, for  $g(x) > 0$  on  $S_X$

$$f(x) = \frac{g(x)}{\int_{x' \in S_X} g(x') df} \quad (4.2)$$

is a density function. Note than normalization is possible only if

$$\sum_{x' \in S_X} g(x') < \infty \text{ or } \int_{x' \in S_X} g(x') df < \infty \quad (4.3)$$

as appropriate. Sometimes we refer to a *normalization constant*  $C$ , that is

$$p(x) = \frac{g(x)}{C} \text{ or } f(x) = \frac{g(x)}{C},$$

where, necessarily,

$$C = \sum_{x' \in S_X} g(x') \text{ or } C = \int_{x' \in S_X} g(x') df,$$

as appropriate.

**Example 4.3** According to statistical genetics certain types of cross breeds occur in the ratios 4:2:2:1. Following (4.1), labeling the outcomes  $S_X = \{1, 2, 3, 4\}$  we set

$$g(1) = 4, \quad g(2) = 2, \quad g(3) = 2, \quad \text{and } g(4) = 1,$$

then the normalization constant is the sum

$$C = \sum_{i=1}^4 g(i) = 4 + 2 + 2 + 1 = 9.$$

which yields PMF

$$p(1) = 4/9, \quad p(2) = 2/9, \quad p(3) = 2/9, \quad \text{and } p(4) = 1/9. \quad \blacksquare$$

**Example 4.4** A *triangle* density is constructed in a piece-wise linear manner as shown in Figure 4.1 (top plot). The base of the triangle is located between  $(0.0, 0.0)$  and  $(1.0, 0.0)$  and the peak is located above  $(0.5, 0.0)$ , assuming value  $f(0.5)$ . Although this is not an exact definition of a single unique function, it suffices, following normalization, to define a single unique density function. Following (4.2) we may select  $g(x)$  to be any function satisfying this definition. Examples are show in Figure 4.1 (bottom plot). Suppose we select:

$$g(x) = \begin{cases} 2x & ; \quad x \in [0, 0.5] \\ 2 - 2x & ; \quad x \in [0.5, 1] \\ 0 & ; \quad \text{ow} \end{cases}$$

which is the solid lined function in Figure 4.1 (bottom plot) (*ow* = ‘otherwise’ denotes all points not yet specified). Then, either by calculus or geometry ( $\text{Area} = 1/2 \times \text{Base} \times \text{Height}$ ), the normalization constant is

$$C = \int_{S_X} g(x) dx = \int_0^{0.5} g(x) dx + \int_{0.5}^{1.0} g(x) dx = 0.5,$$

so that the normalized density function is

$$f(x) = 2g(x) = \begin{cases} 4x & ; \quad x \in [0, 0.5] \\ 4 - 4x & ; \quad x \in [0.5, 1] \\ 0 & ; \quad \text{ow} \end{cases}. \quad (4.4)$$

First note that we would attain exactly the same density starting from any of the functions shown in Figure 4.1 (bottom plot). Also note that we have been somewhat flexible with our conventions. The density (4.4) formally has sample space  $S_X = (-\infty, \infty)$ , whereas the support is  $\mathcal{S}_X = (0.0, 1.0)$ . It would not be incorrect to have defined both the sample space and support to be  $(0.0, 1.0)$ . ■

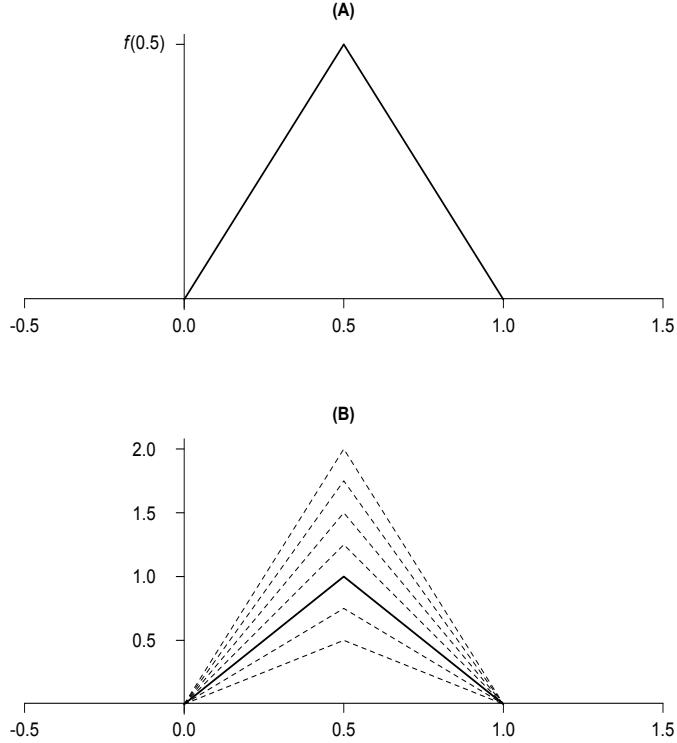


Figure 4.1: Triangle density for Example 4.4. The top plot (A) shows the actual triangle density  $f(x)$ , while the bottom plot (B) shows examples of functions  $g(x)$  which, when normalized, will yield the triangle density  $f(x)$ .

### 4.3 Cumulative Distribution Functions

An alternative form of representing a RV is introduced in the next definition:

**Definition 4.3** The of a random variable  $X$  is defined as

$$F_X(x) = P(X \leq x), \quad x \in (-\infty, \infty). \quad ■$$

This definition is the same for any type of RV, but the evaluation method depends on the type. If  $X$  is a discrete random variable with support  $\mathcal{S}_X$ , the CDF may be defined as

$$F_X(x) = P(X \leq x) = \sum_{u \in \mathcal{S}_X : u \leq x} P(X = u). \quad (4.5)$$

If, for example,  $X$  is a positive integer, we have:

$$F_X(k) = \sum_{i=1}^k P(X = i), \quad k = 1, 2, \dots$$

The CDF of a continuous random variable  $X$  is calculated by the integral

$$F_X(x) = P(X \leq x) = \int_{-\infty}^x f_X(u) du. \quad (4.6)$$

It is sometimes more convenient to consider the *tail probability*  $P(X > x) = 1 - P(X \leq x)$ . This is denoted

$$\bar{F}_X(x) = P(X > x) = 1 - F_X(x).$$

For *any* type of random variable, the definition  $F_X(x) = P(X \leq x)$  always holds, and the CDF is defined for all  $x \in (-\infty, \infty)$ , even when the support  $S_X$  is a strictly smaller set.

Taking the derivative of both sides of (4.6) yields the following important relationship:

**Theorem 4.1** For a continuous RV  $X$ , the following relationship holds,

$$\frac{d}{dx} F_X(x) = f_X(x), \quad (4.7)$$

wherever the derivative exists. ■

**Example 4.5** Consider the distribution of Example 4.9. We had support  $S_X = \{0, 1, 2, 3\}$ , and the CDF (4.5) is straightforward to calculate on these

points:

$$\begin{aligned}
 F_X(0) &= P(X \leq 0) \\
 &= P(X = 0) \\
 &= 1/8, \\
 F_X(1) &= P(X \leq 1) \\
 &= P(X \in \{0, 1\}) \\
 &= P(X = 0) + P(X = 1) \\
 &= 1/8 + 3/8 \\
 &= 4/8, \\
 F_X(2) &= P(X \leq 2) \\
 &= P(X \in \{0, 1, 2\}) \\
 &= P(X = 0) + P(X = 1) + P(X = 2), \\
 &= 1/8 + 3/8 + 3/8 \\
 &= 7/8 \\
 F_X(3) &= P(X \leq 3) \\
 &= P(X \in \{0, 1, 2, 3\}) \\
 &= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) \\
 &= 1/8 + 3/8 + 3/8 + 1/8 \\
 &= 8/8 \\
 &= 1.
 \end{aligned}$$

Of course, the CDF may be defined on the entire real line, irrespective of the support or sample space. Nothing prevents us from calculating  $F_X(1.5) = P(X \leq 1.5)$ . Since the support of  $X$  is  $\mathcal{S}_X = \{0, 1, 2, 3\}$  then  $X \leq 1.5$  is equivalent to  $X \in \{0, 1\}$ , or  $X \leq 1$ , so that  $F_X(1.5) = F_X(1) = 4/8$ . Similary, for any number  $x < 0$ , we can say that  $F_X(x) = P(X \leq x) = 0$ , and for any number  $x > 3$  we have  $F_X(x) = P(X \leq x) = 1$ . Thus, the CDF for a discrete random variable is a nondecreasing *step function*. The CDF for the current example is shown in Figure 4.2. ■

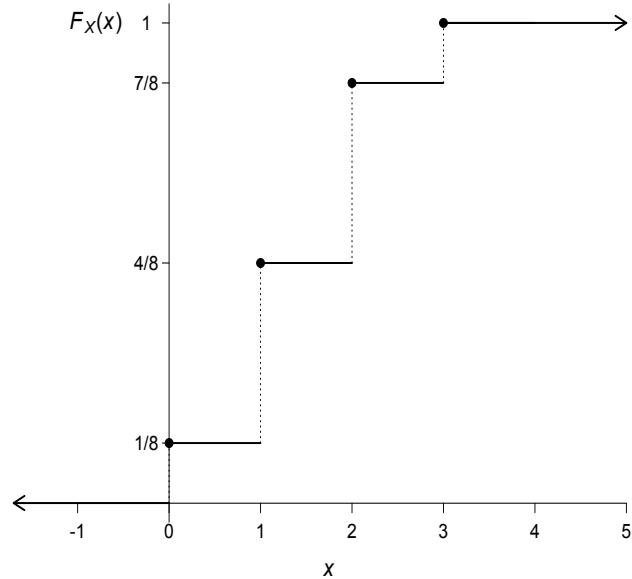


Figure 4.2: CDF for Example 4.5. Note that the function  $F_X(x)$  extends to  $\pm\infty$ .

**Example 4.6** Consider the triangle density of Example 4.4. Using (4.6) to evaluate the CDF will have to be in a piecewise manner, in particular, 4 pieces, as follows.

**Range**  $x \in (-\infty, 0)$ . First, recall that the CDF is defined on the entire real line  $(-\infty, \infty)$ . However, for a RV  $X$  with density (4.4) we have  $P(X \leq 0) = 0$ , and so  $P(X \leq x) = 0$  for any  $x < 0$ . This means

$$F_X(x) = 0 \text{ for } x \in (-\infty, 0).$$

**Range**  $x \in [0, 0.5)$ . In this range using evaluation method (4.6) gives

$$\begin{aligned} F_X(x) &= \int_{-\infty}^x f(u)du \\ &= \int_{-\infty}^0 f(u)du + \int_0^x f(u)du \\ &= 0 + \int_0^x 4udu \\ &= 2u^2 \Big|_0^x \\ &= 2x^2 \text{ for } x \in [0, 0.5). \end{aligned}$$

**Range**  $x \in [0.5, 1.0]$ . In this range using evaluation method (4.6) gives

$$\begin{aligned}
F_X(x) &= \int_{-\infty}^x f(u)du \\
&= \int_{-\infty}^0 f(u)du + \int_0^{0.5} f(u)du + \int_{0.5}^x f(u)du \\
&= 0 + 2(0.5)^2 + \int_{0.5}^x 4 - 4udu \\
&= 2(0.5)^2 + (4u - 2u^2) \Big|_{0.5}^x \\
&= 0.5 + (4x - 2x^2) - (4(0.5) - 2(0.5)^2) \\
&= -2x^2 + 4x - 1 \text{ for } x \in [0.5, 1.0].
\end{aligned}$$

**Range**  $x \in (1.0, \infty)$ . For a RV  $X$  with density (4.4) we have  $P(X \leq 1) = 1$ , and so  $P(X \leq x) = 1$  for any  $x > 1.0$ . This means

$$F_X(x) = 1 \text{ for } x \in (1.0, \infty).$$

To express  $F_X$  analytically requires a piecewise definition:

$$F_X(x) = \begin{cases} 0 & ; \quad x \in (-\infty, 0) \\ 2x^2 & ; \quad x \in [0, 0.5) \\ -2x^2 + 4x - 1 & ; \quad x \in [0.5, 1.0] \\ 1 & ; \quad x \in (1.0, \infty) \end{cases}. \quad (4.8)$$

The CDF for the current example is shown in Figure 4.3. ■

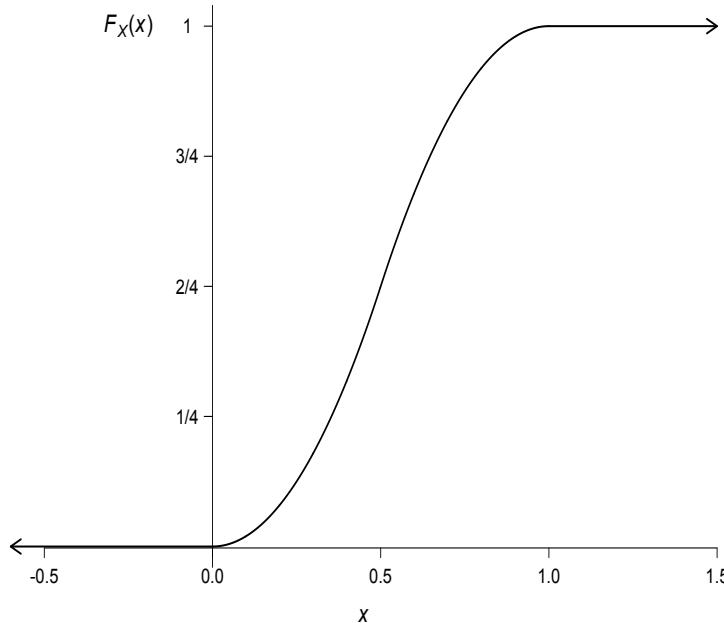


Figure 4.3: CDF for Example 4.6. Note that the function  $F_X(x)$  extends to  $\pm\infty$ .

## 4.4 Quantiles and Percentiles

Suppose it is reported that the 5-year recurrence rate of large cell hepatocellular carcinoma (a type of liver cancer) after curative resection is 61.5%. We can formulate this as a probability statement. Let  $X$  be a RV which models the time until recurrence after curative resection. Then

$$P(X \leq 5) = 0.615,$$

that is, for 61.5% of patients the recurrence time  $X$  is less than 5. We say that  $q = 5$  is the  $p = 0.615$  quantile, or equivalently, the 61.5*th* percentile.

The notion of a quantile is quite intuitive. For example, if we write a standardized test (such as the GRE), we are interested in interpreting our score as a quantile, or percentile. If a score  $x$  is the 0.75-quantile (75*th* percentile), then  $x$  is higher than 75% of all scores, and less than 25% of all scores. However, there are some unintuitive mathematical details which must be understood.

We start with a formal definition (you may wish to review the definition of the inverse function in Appendix B.1 and the definition of the infimum in Appendix B.2):

**Definition 4.4** Suppose  $X$  is any RV. Then  $q$  is a *quantile* ( $p$ -quantile) of  $X$  if

$$P(X < q) \leq p \text{ and } P(X > q) \leq 1 - p. \quad (4.9)$$

A  $p$ -quantile is, equivalently, a  $p \times 100$ th *percentile*.

The *quantile function* of  $X$  is

$$Q(p) = \inf\{x \in \mathbb{R} : P(X \leq x) \geq p\}. \quad (4.10)$$

If the CDF  $F_X(x)$  is continuous and strictly increasing on support  $\mathcal{S}_X$ , then it possesses an inverse  $F_X^{-1}$  which maps  $[0, 1]$  to  $\mathcal{S}_X$ , and we have

$$Q(p) = F_X^{-1}(p). \quad (4.11)$$

In this case the  $p$ -quantile is unique. ■

Some attention needs to be paid to the wording in Definition 4.4. The value of the quantile function  $Q(p)$  is uniquely defined, but the  $p$ -quantile need not be. When it is, it will be equal to  $Q(p)$ . To take a simple example, suppose a RV  $X$  equals 0 or 1, each with probability 1/2. Then consider the inequalities of (4.9), and suppose  $0 < q < 1$ . Then

$$P(X < q) \leq 1/2 \text{ and } P(X > q) \leq 1/2,$$

so that any number  $q$  strictly between 0 and 1 is a 0.5-quantile. If we evaluate the quantile function (4.10) at  $p = 0.5$  we get

$$Q(0.5) = \inf\{x \in \mathbb{R} : P(X \leq x) \geq 0.5\} = \inf[0, \infty) = 0.$$

Of course, quantiles are usually of more interest for either continuous RVs, or discrete RVs with large support.

**Example 4.7** Consider the triangle density of Example 4.4. The CDF was derived in Example 4.6. The conditions given in Definition 4.4 under which (4.11) holds are satisfied here. Figure 4.4 shows a plot of this CDF, and shows the relationship implied by the CDF and quantile function

$$p' = F_X(q') \text{ and } q' = Q(p') = F_X^{-1}(p').$$

If  $p' = 1/8$ , then from the expression for the CDF given in (4.8) we have, assuming  $p' < 0.5$ ,

$$p' = 2(q')^2 \text{ or } q' = (p'/2)^{1/2},$$

so that

$$q' = (0.125/2)^{1/2} = 0.25$$

is the 0.125-quantile. ■

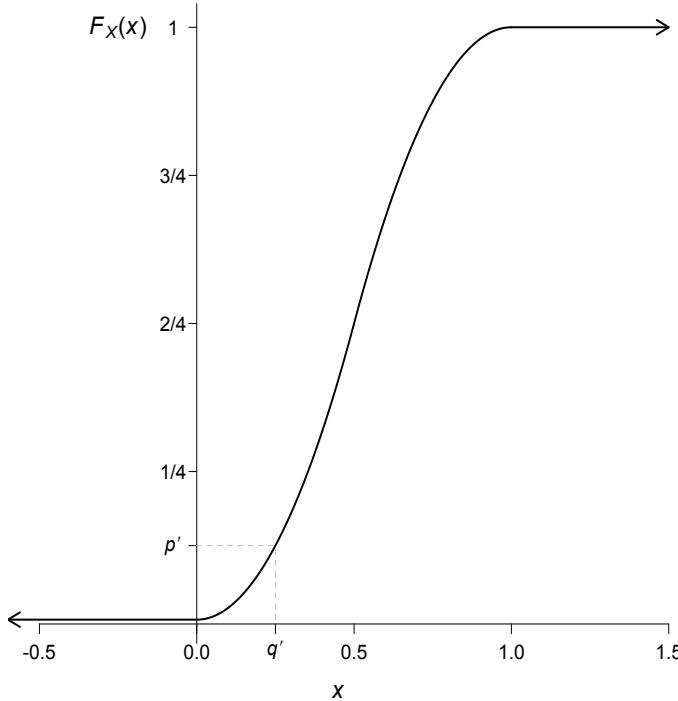


Figure 4.4: CDF and quantile for Example 4.7. Note that the function  $F_X(x)$  extends to  $\pm\infty$ .

It is certainly more convenient when the  $p$ -quantile is unique and equal to the quantile function  $Q(p)$ . However, this convenient relationship will not hold under two conditions. First, suppose  $P(X = q') = p' > 0$ . In this case,  $q'$  will be a  $p$ -quantile for multiple values of  $p$ . Second, suppose  $P(X \in (a, b)) = 0$  for some nonempty open interval  $(a, b)$ . In this case, the value

$$p' = F_X(q')$$

is constant for all  $q' \in (a, b)$ , and all  $q'$  in this interval satisfy the definition of a  $p'$ -quantile. These cases are illustrated in the next example.

**Example 4.8** Consider the CDF illustrated in Figure 4.5. First, note that  $P(X = q_3) = p_3 - p_2 > 0$ . We also have

$$\begin{aligned} P(X < q_3) &= p_2, \text{ and} \\ P(X > q_3) &= 1 - p_3. \end{aligned}$$

It is easily verified that inequalities (4.9) of Definition 4.4 are satisfied for  $q = q_3$  any  $p \in [p_2, p_3]$ , that is,  $q_3$  is a  $p$ -quantile for any  $p \in [p_2, p_3]$ .

Next, we note that  $P(X \in (q_1, q_2)) = 0$ . It may also be verified that inequalities (4.9) of Definition 4.4 are satisfied for  $p = p_1$  and any  $q \in [q_1, q_2]$ . This means any  $q \in [q_1, q_2]$  is a  $p_1$ -quantile. ■

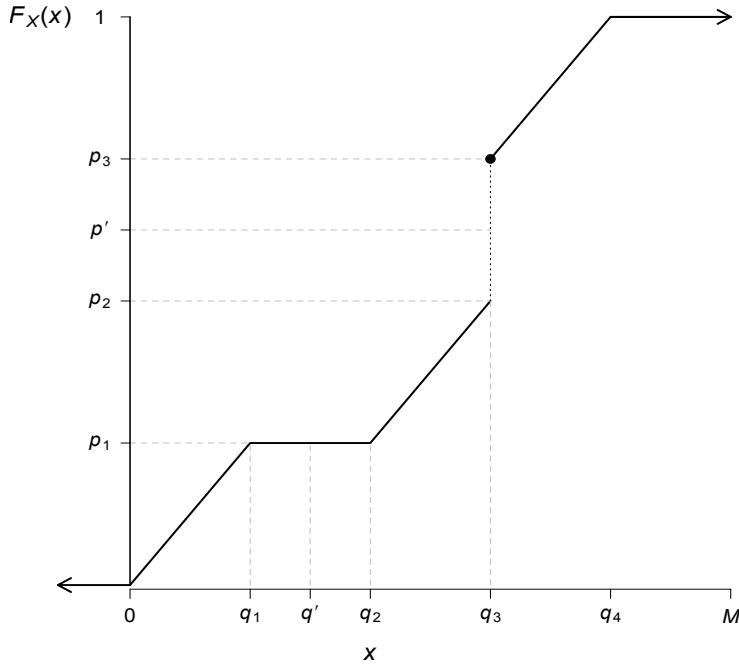


Figure 4.5: CDF for Example 4.6. Note that the function  $F_X(x)$  extends to  $\pm\infty$ .

#### 4.4.1 Critical Values

A closely related idea to the quantile is the *critical value*. Suppose  $X$  has some distribution defined by  $F_X$ . Then the  $\alpha$ -critical value  $x_\alpha$  is a number satisfying

$$P(X \geq x_\alpha) = \alpha.$$

If  $X$  is a continuous *RV* we have  $x_\alpha = Q(1-\alpha)$ . Usually, but not always, the intention is that  $\alpha$  is small, so that  $x_\alpha$  would be a relatively large observation from this distribution.

## 4.5 Expected Value

Suppose we have an integer valued random variable  $X$  with PMF  $p_i = P(X = i)$ ,  $i = 0, 1, \dots, N$ . Suppose we then observe  $n$  random variables from this distribution, say  $X_1, X_2, \dots, X_n$ . The average observation would

be

$$\begin{aligned}\bar{X}_n &= \frac{X_1 + X_2 + \dots + X_n}{n} \\ &= \frac{n_0 \times 0 + n_1 \times 1 + n_2 \times 2 + \dots + n_N \times N}{n} \\ &= \frac{n_0}{n} \times 0 + \frac{n_1}{n} \times 1 + \frac{n_2}{n} \times 2 + \dots + \frac{n_N}{n} \times N,\end{aligned}$$

where  $n_i$  is the number of RVs which equal  $i$ . Of course, it is reasonable to expect

$$\frac{n_i}{n} \approx p_i$$

for any  $i$ . In fact, one of the most important facts of probability theory is the *strong law of large numbers* (SLLN) which says, among other things, that this approximation becomes increasingly accurate as  $n$  increases (formally, with probability one,  $n_i/n$  converges to  $p_i$  as  $n$  approaches  $\infty$ ).

This then forces a conclusion about  $\bar{X}_n$ , namely that

$$\bar{X}_n \approx \sum_{i=0}^N ip_i = E[X],$$

and that this approximation becomes increasingly accurate as  $n$  increases (formally, with probability one,  $\bar{X}_n$  converges to  $E[X]$  as  $n$  approaches  $\infty$ ). This is another consequence of the SLLN.

We then have the following definition:

**Definition 4.5** The quantity  $E[X]$  is called the *expected value* of  $X$ , and represents a theoretical average of an infinitely large sample. The notation

$$\mu = \mu_X = E[X]$$

is frequently used.

For discrete random variables, the expected value is calculated by

$$E[X] = \sum_{x \in S} xP(X = x), \quad (4.12)$$

that is, for every possible outcome in  $S_X$  we contribute  $xP(X = x)$  to the sum.

The expected value for continuous random variables  $X$  with density  $f_X$  is evaluated by the integral

$$E[X] = \int_{-\infty}^{\infty} xf_X(x)dx. \quad (4.13)$$

We lose no generality in defining  $(-\infty, \infty)$  to be the domain of integration even when the sample space or support is strictly smaller. This is because any contribution to the integral outside the support must be exactly 0.

Note that  $E[X]$  is commonly referred to as the *mean*, or *population mean*. ■

**Example 4.9** We toss a fair coin ( $H$  or  $T$ ) independently until a head  $H$  appears. Let  $X$  be the total number of tosses required. We will find the PMF and expected value  $E[X]$  for  $X$ .

First, note that  $X = 1$  if and only if the first toss is  $H$ , which has probability  $1/2$ . This means that  $P(X = 1) = 1/2$ . Similarly,  $X = 2$  can only occur if the first toss is  $T$  and the second is  $H$ . By independence this has probability  $1/2 \times 1/2 = 1/4$ . By the same logic,  $X = k$  if and only if the first  $k - 1$  tosses were  $T$  and the  $k$ th toss was  $H$ . This event specifies exactly one outcome for each of  $k$  tosses, so we conclude  $P(X = k) = (1/2)^k$ . This defines a RV with support  $\mathcal{S} = \{1, 2, \dots\}$  and PMF

$$p_X(k) = (1/2)^k, \quad k = 1, 2, \dots$$

We can easily verify that  $p_X$  is a proper PMF by evaluation the geometric series

$$1/2 + (1/2)^2 + (1/2)^3 + \dots = \frac{1}{1 - 1/2} - 1 = 1,$$

(see Section B.5). Here we have an example of a RV with discrete but infinite support.

To evaluate  $E[X]$  we use (4.12) to obtain expression:

$$\begin{aligned} E[X] &= \sum_{i=1}^{\infty} i(1/2)^i \\ &= 1(1/2)^1 + 2(1/2)^2 + \dots \\ &= (1/2) \times (1(1/2)^0 + 2(1/2)^1 + \dots) \\ &= (1/2) \frac{1}{(1 - 1/2)^2} \\ &= 2 \end{aligned}$$

(compare with the forms for the geometric series in Section B.5). Thus, on average, two coin tosses are required to produce the first head.

This random variable is a special case of the *geometric random variable* which we will consider in Section 4.10.4 ■

**Example 4.10** Returning to the triangle density of Example 4.4, we might conjecture on the basis of symmetry that the expected value for density function (4.4) will be  $E[X] = 0.5$ . This is correct, and this type of argument can be made mathematically rigorous. Here, we will simply verify this using

the evaluation method (4.13):

$$\begin{aligned}
E[X] &= \int_{-\infty}^{\infty} uf(u)du \\
&= \int_0^{0.5} u \times 4udu + \int_{0.5}^{1.0} u \times (4 - 4u)du \\
&= \frac{4}{3}u^3 \Big|_0^{0.5} + \left(2u^2 - \frac{4}{3}u^3\right) \Big|_{0.5}^{1.0} \\
&= \frac{4}{3}(0.5)^3 + \left(2(1.0)^2 - \frac{4}{3}(1.0)^3\right) - \left(2(0.5)^2 - \frac{4}{3}(0.5)^3\right) \\
&= 1/2.
\end{aligned}$$
■

## 4.6 Functions of Random Variables

Given a random variable  $X$  and a function  $f$  we easily obtain a new random variable  $Y = g(X)$ . This operation is often referred to as a *transformation*. A few questions arise. First, how can the distribution of  $Y$  (as PMF, density or CDF) be deduced from that of  $X$ ? Second, if we are interested only in, for example,  $E[Y]$ , do we need to determine its complete distribution?

As a general rule, distribution  $P_Y$  of  $Y = g(X)$  can be directly obtained from the distribution  $P_X$  of  $X$  using the idea of the *preimage* (see Section B.1):

$$g^{-1}(E) = \{x \in \mathbb{R} : g(x) \in E\}$$

which means that

$$\{Y \in E\} = \{X \in g^{-1}(E)\},$$

so that

$$P_Y(E) = P_X(g^{-1}(E)).$$

Note that the use of the preimage does not require that the function  $g$  actually have an inverse ((see Section B.1)). An important example is given next.

### 4.6.1 CDF Method

Suppose  $Y = g(X)$  and  $g$  is a strictly increasing function. Then  $g$  possesses a strictly increasing inverse function  $g^{-1}$ . In general, if  $h$  is any increasing function, and  $x \leq y$  then we also have  $h(x) \leq h(y)$ . We may therefore write

$$\{Y \leq y\} = \{g(X) \leq y\} = \{g^{-1}(g(X)) \leq g^{-1}(y)\} = \{X \leq g^{-1}(y)\}. \quad (4.14)$$

We have the following result:

**Theorem 4.2** Suppose  $X$  is a discrete or continuous RV with CDF  $F_X$ . Let  $g$  be a strictly increasing function. Then the RV  $Y = g(X)$  possesses CDF

$$F_Y(y) = F_X(g^{-1}(y)), \quad (4.15)$$

noting that the inverse  $g^{-1}$  must be well defined. ■

**Proof.** Equation (4.15) follows from the equality

$$\{Y \leq y\} = \{X \leq g^{-1}(y)\}$$

derived in Equation (4.14).  $\square$

**Example 4.11** Recall the triangle density of 4.4. The CDF is given in Example 4.6. Suppose RV  $X$  possesses this density, and we wish to evaluate the density of  $Y = X^2$ . We can proceed by first evaluating the CDF  $F_Y$  then taking the derivative, as shown in Equation 4.7.

In this case we have the transformation  $Y = g(X)$ , with  $g(x) = x^2$ . Although  $g$  is not increasing on  $\mathbb{R}$ , it is increasing on the support  $\mathcal{S} = [0, 1]$  of  $X$ . This suffices for 4.14 to hold where needed, and so we may use equation 4.15. This is easily done, since all that is needed is to replace  $x$  in 4.6 with  $\sqrt{y}$ :

$$F_Y(y) = \begin{cases} 0 & ; y \in (-\infty, 0) \\ 2y & ; y \in [0, 0.25) \\ -2y + 4\sqrt{y} - 1 & ; y \in [0.25, 1.0] \\ 1 & ; y \in (1.0, \infty) \end{cases}$$

noting the adjustments made to the segment ranges, for example

$$\sqrt{y} \in [0, 0.5] \iff \sqrt{y} \in [0, 0.25].$$

The density  $f_Y$  follows directly as the derivative of  $F_Y$ :

$$f_Y(y) = \begin{cases} 0 & ; y \in (-\infty, 0) \\ 2 & ; y \in [0, 0.25) \\ \frac{2}{\sqrt{y}} - 2 & ; y \in [0.25, 1.0] \\ 1 & ; y \in (1.0, \infty) \end{cases}.$$

The densities are shown in Figure 4.6.  $\blacksquare$

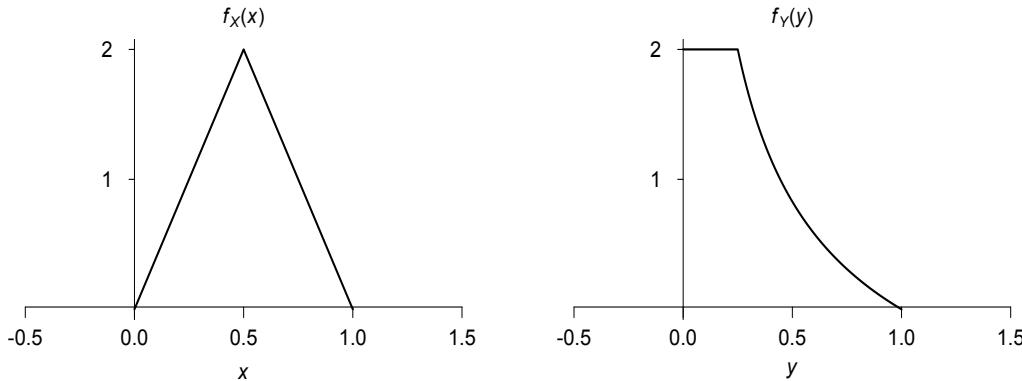


Figure 4.6: Densities of  $X$  (triangle density) and  $Y = X^2$  for Example 4.11.

### 4.6.2 One-to-one Transformations

Suppose the transformation  $Y = g(X)$  is one-to-one. Then the transformation possesses an inverse  $g^{-1}$  between the support of  $X$  and the support of  $Y$ , which justifies its use. If  $X$  is discrete, the PMFs  $p_X$  and  $p_Y$  can be related by direct substitution:

$$p_Y(y) = P(Y = y) = P(g(X) = y) = P(X = g^{-1}(y)) = p_X(g^{-1}(y)).$$

If  $X$  is continuous, a similar, but more complex method is available. We'll return temporarily to the assumption that  $g$  is increasing. If we differentiate both sides of (4.14) using the chain rule we have

$$\frac{d}{dy} F_Y(y) = \frac{dg^{-1}(y)}{dy} \frac{d}{dy} F_X(g^{-1}(y)).$$

Recalling that the density  $f$  is the derivative of the CDF  $F$  (Equation 4.7) we have

$$f_Y(y) = \left[ \frac{dg^{-1}(y)}{dy} \right] f_X(g^{-1}(y)) = \left[ \frac{dg(x)}{dx} \Big|_{x=x(y)} \right]^{-1} f_X(g^{-1}(y)).$$

However, it can be shown that this method does not depend on  $g$  being increasing. As long as it is a one-to-one transform, we have the transformation rule

$$f_Y(y) = \left| \frac{dg^{-1}(y)}{dy} \right| f_X(g^{-1}(y)) = \left| \frac{dg(y)}{dy} \right|_{x=x(y)}^{-1} f_X(g^{-1}(y)),$$

replacing the derivative of the transformation with its absolute value.

We summarize this section with the following theorem

**Theorem 4.3** Suppose  $X$  is a RV, and  $g$  is a function possessing inverse  $g^{-1}$ . Define RV  $Y = g(X)$ . If  $X$  is discrete with PMF  $p_X$  then  $Y$  is discrete, and possesses PMF

$$p_Y(y) = P(Y = y) = P(g(X) = y) = P(X = g^{-1}(y)) = p_X(g^{-1}(y)).$$

If  $X$  is continuous with density  $f_X$  then  $Y$  is continuous, and possesses density

$$f_Y(y) = \left| \frac{dg^{-1}(y)}{dy} \right| f_X(g^{-1}(y)).$$
■

### 4.6.3 Expected Values of Functions of Random Variables

In the previous Sections 4.6.1 and 4.6.2 we considered methods for determining the distribution of a transformed random variable  $Y = g(X)$ . If we are interested in evaluating  $E[Y]$  we can first determine the PMF or density of  $Y$ , then use the appropriate method from Definition 4.5. However, we may also evaluate  $E[Y]$  directly from the distribution of  $X$ , as shown in the following theorem.

**Theorem 4.4** Suppose  $X$  is a RV, and  $g$  is a function. Define RV  $Y = g(X)$ . If  $X$  is discrete with PMF  $p_X$  then

$$E[Y] = \sum_{x \in \mathcal{S}_X} g(x)p_X(x).$$

If  $X$  is continuous with density  $f_X$  then

$$E[Y] = \int_{-\infty}^{\infty} g(x)f_X(x)dx.$$
■

A number of important results follow directly from Theorem 4.4.

**Theorem 4.5** For any constant  $c$

$$E[c] = c.$$

Suppose  $X$  is a RV, and  $a, b$  are two constants. Then

$$E[aX + b] = aE[X] + b.$$

If  $g_1, g_2$  are two functions, then

$$E[g_1(X) + g_2(X)] = E[g_1(X)] + E[g_2(X)].$$
■

Note that Theorem 4.5 recognizes the possibility that a random variable  $X$  is constant. Formally, this means that there is a constant  $c$  such that

$$P(X = c) = 1. \tag{4.16}$$

This is perfectly valid.

We define an important class of expected values.

**Definition 4.6** The  $k$ th moment of random variable is defined to be the quantity  $E[X^k]$ .

■

## 4.7 Variance

The expected value  $E[X]$  represents a type of average of  $X$ . The *variance* measures the tendency of  $X$  to deviate from  $E[X]$ , and is introduced in the following definition.

**Definition 4.7** The *variance*  $\text{var}[X]$  of a random variable is defined by

$$\text{var}[X] = E[(X - E[X])^2].$$
■

The notation

$$\sigma^2 = \sigma_X^2 = \text{var}[X]$$

is frequently used. The *standard deviation*  $\sigma$  is the square root of the variance, conventionally written

$$\sigma = \sigma_X = \sqrt{\text{var}[X]}.$$

Formally,  $\sqrt{(X - E[X])^2} = |(X - E[X])|$ , but it is generally not the case that the standard deviation  $\sigma$  is equal to  $E[|(X - E[X])|]$ . The one exception is the constant RV  $X \equiv c$  (see Equation (4.16)). In this, by Theorem 4.5 we have

$$\begin{aligned} E[X] &= E[c] = c, \text{ and} \\ \text{var}[X] &= E[c - c] \\ &= c - c \\ &= 0 \end{aligned} \tag{4.17}$$

That the variance of a constant is 0 is reasonable, since it measures the tendency of a RV to deviate from its mean.

An alternative form for the variance follows from Theorem 4.5.

**Theorem 4.6** The variance of a random variable is equivalent to

$$\text{var}[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2. \quad \blacksquare$$

**Proof.** First note that

$$(X - E[X])^2 = X^2 - 2XE[X] + E[X]^2.$$

then, using Theorem 4.5 we have

$$\begin{aligned} E[(X - E[X])^2] &= E[X^2 - 2XE[X] + E[X]^2] \\ &= E[X^2] + E[-2XE[X]] + E[E[X]^2] \\ &= E[X^2] - 2E[X]E[X] + E[X]^2 \\ &= E[X^2] - 2E[X]^2 + E[X]^2 \\ &= E[X^2] - 2E[X]^2, \end{aligned}$$

which completes the proof.  $\square$

#### 4.7.1 Variance of Linear Transformations

Suppose  $X$  is a RV and  $a, b$  are two constants. We sometimes need to construct a new RV as a *linear transformation*:

$$Y = aX + b.$$

Then  $\text{var}[Y]$  is easily related to  $\text{var}[X]$ .

**Theorem 4.7** If  $X$  is a RV and  $a, b$  are constants, and if  $Y = aX + b$  then

$$\text{var}[Y] = a^2\text{var}[X]. \tag{4.18}$$

In particular,  $\text{var}[Y]$  does not depend on  $b$ .  $\blacksquare$

**Proof.** From Theorem 4.5 we have  $E[Y] = aE[X] + b$ . To prove (4.18) note that

$$\begin{aligned}\text{var}[Y] &= E[(aX + b - E[aX + b])^2] \\ &= E[(aX + b - aE[X] - b)^2] \\ &= E[(aX - aE[X])^2] \\ &= E[a^2(X - aE[X])^2] \\ &= a^2E[(X - E[X])^2] \\ &= a^2\text{var}[X].\end{aligned}$$

Note that we have made use of Theorem 4.5 a second time in the preceding argument.  $\square$

## 4.8 Random Variables and Independence

We should have no objection to identifying multiple random variables with a single random outcome. For example, if we toss 2 labelled dice, we may assign the respective outcomes to random variables  $X_1$  and  $X_2$ . The notion of independence extends to random variables in a natural way, since we may always test the independence of events of the form  $A_1 = \{X_1 \in E_1\}$ ,  $A_2 = \{X_2 \in E_2\}$  for any real valued subsets  $E_1, E_2$  using the product rule of Definition 2.8. This leads to the following definition:

**Definition 4.8** A collection of random variables  $X_1, X_2, \dots, X_n$  is *independent* if for any collection of real valued sets

$$E_1, E_2, \dots, E_n$$

the events

$$A_i = \{X_i \in E_i\}, \quad i = 1, 2, \dots, n$$

is independent according to Definition 2.8.

In particular two RVs  $X, Y$  are independent if

$$P(X_1 \in E_1, X_2 \in E_2) = P(X_1 \in E_1)P(X_2 \in E_2)$$

for all real-valued subsets  $E_1, E_2$ . In this case we write  $X_1 \perp X_2$ .  $\blacksquare$

Note that the independence of a collection of random variables requires the independence of any subcollection.

We have the following important definition:

**Definition 4.9** A collection of random variables  $X_1, X_2, \dots, X_n$  is an *iid sample* if they are independent and have the same distribution (*iid* mean *independent and identically distributed*).  $\blacksquare$

**Example 4.12** Suppose  $X_1$  and  $X_2$  are the outcomes of two labelled dice, and that each outcome  $(i, j)$ ,  $i = 1, \dots, 6$ ,  $j = 1, \dots, 6$  are equally likely, and so each have probability  $1/36$ . It is reasonable to suppose that  $X_1 \perp X_2$ . However, to test Definition 4.8 we need to consider all pairs of outcome events  $E_1, E_2$ , which are each subjects of  $\{1, \dots, 6\}$ . Suppose  $n_i = |E_i|$ ,  $i = 1, 2$ . Then we have the probabilities

$$\begin{aligned} P(X_1 \in E_1) &= n_1/6 \\ P(X_2 \in E_2) &= n_2/6 \\ P(X_1 \in E_1, X_2 \in E_2) &= n_1 n_2 / 36 \end{aligned}$$

the last probability following from the fact that the event has  $n_1 n_2$  outcomes (this type of problem will be considered in more detail in Chapter ). It is easily verified that the product rule holds,

$$P(X_1 \in E_1, X_2 \in E_2) = \frac{n_1 n_2}{36} = (n_1/6) \times (n_2/6) = P(X_1 \in E_1)P(X_2 \in E_2),$$

so that  $X_1 \perp X_2$ . ■

A counter-example to independence is not hard to construct.

**Example 4.13** Continuing from Example 4.12, suppose we define two new random variables

$$Y_1 = \min(X_1, X_2), \quad Y_2 = \max(X_1, X_2).$$

Often, independence is easier to disprove than to prove, since we only require one violation of Definition 4.8. Following the notational conventions of Example 4.12 set  $E_1 = \{4\}$  and  $E_2 = \{3\}$ . By construction  $X_1 \leq X_2$  and so

$$P(X_1 \in E_1, X_2 \in E_2) = 0.$$

All that remains is to note that  $P(X_1 \in E_1) > 0$  and  $P(X_2 \in E_2) > 0$ , so that

$$0 = P(X_1 \in E_1, X_2 \in E_2) \neq P(X_1 \in E_1)P(X_2 \in E_2) > 0,$$

and we conclude that  $X_1$  and  $X_2$  are not independent. ■

The notion of independence extends naturally to expectations.

**Theorem 4.8** If  $X_1, X_2, \dots, X_n$  are independent random variables, and  $g_i(x)$ ,  $i = 1, 2, \dots, n$  are  $n$  functions, then

$$E \left[ \prod_{i=1}^n g_i(X_i) \right] = \prod_{i=1}^n E[g_i(X_i)]$$

assuming all expectations are finite. ■

### 4.8.1 Covariance and Correlation

The *covariance* and *correlation* are commonly used measures of association between two RVs.

**Definition 4.10** Suppose  $X, Y$  are two RVs with means  $\mu_X, \mu_Y$  and variances  $\sigma_X^2, \sigma_Y^2$ . The *covariance* of  $X$  and  $Y$  is defined as

$$\text{cov}[X, Y] = \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)].$$

The *correlation* of  $X$  and  $Y$  is defined as

$$\text{cor}[X, Y] = \rho = \rho_{XY} = \frac{\text{cov}[X, Y]}{\sigma_X \sigma_Y}.$$

Note that the correlation is defined only if both variances are positive, whereas this restriction does not hold for the covariance. ■

The following theorem summarizes some important facts concerning covariances and correlations

**Theorem 4.9** Suppose  $X, Y$  are two RVs with means  $\mu_X, \mu_Y$  and variances  $\sigma_X^2, \sigma_Y^2$ . The following statements hold.

- (i) If  $X$  or  $Y$  is constant, then  $\text{cov}[X, Y] = 0$ , and  $\text{cor}[X, Y]$  is undefined.
- (ii) If  $X \perp Y$  then  $\text{cov}[X, Y] = 0$  and, when defined,  $\text{cor}[X, Y] = 0$ .
- (iii) We always have

$$-1 \leq \text{cor}[X, Y] \leq 1. \quad (4.19)$$

- (iv)  $\text{cov}[X, X] = \text{var}[X]$ .
- (v)  $\text{cov}[X, Y] = \text{cov}[Y, X]$ . ■

*Proof.* We consider each statement in sequence.

(i) If, say,  $X \equiv c$  is a constant, then  $X - \mu_X = 0$  is also a constant, so that by Theorem 4.5 we have

$$E[(X - \mu_X)(Y - \mu_Y)] = E[0 \times (Y - \mu_Y)] = 0 \times E[(Y - \mu_Y)] = 0$$

The correlation is undefined because  $\sigma_X^2 = 0$  (see Equation (4.17)). The proof is identical if  $Y$  is a constant.

- (ii) We have from Theorem 4.5

$$E[X - \mu_X] = E[X] - \mu_X = \mu_X - \mu_X = 0 \text{ and similarly } E[Y - \mu_Y] = 0.$$

By Theorem 4.8, if  $X \perp Y$  we have

$$\begin{aligned} \text{cov}[X, Y] &= E[(X - \mu_X)(Y - \mu_Y)] \\ &= E[(X - \mu_X)] E[(Y - \mu_Y)] \\ &= 0 \times 0 \\ &= 0. \end{aligned}$$

(iii) This statement follows from the *Cauchy-Schwartz inequality* which, specialized to probability theory states that for any two random variables  $X, Y$  we always have

$$E[XY] \leq \left( E[X^2] E[Y^2] \right)^{1/2}. \quad (4.20)$$

Then (4.19) follows by substituting  $(X - \mu_X)$  and  $(Y - \mu_Y)$  for  $X$  and  $Y$  in (4.20).

(iv) We have, directly,

$$\text{cov}[X, X] = E[(X - \mu_X)(X - \mu_X)] = E[(X - \mu_X)^2] = \text{var}[X].$$

(v) The symmetry of  $\text{cov}[X, Y]$  follows directly from the definition.  $\square$

It is important to note that while independence implies zero covariance, the converse is not true. A counter-example follows.

**Example 4.14** We can model the position of a dart on a dartboard as a random experiment. If the player is not particularly skilled, the dart might be equally likely to occupy any position. Suppose the dartboard is represented by circle  $A$ . Let  $E$  be any region in  $A$ . Formally, under this model the position of the dart, denoted  $(X, Y)$  is *uniformly distributed* on  $A$ , which is equivalent to the evaluation method

$$P((X, Y) \in E) = \frac{\text{area}(E)}{\text{area}(A)}. \quad (4.21)$$

we'll take  $A$  to be a circle in Cartesian coordinates with center  $(0, 0)$  and unit radius, so that we have  $X^2 + Y^2 \leq 1$ .

In general, the analysis of joint distributions is often technically challenging. However, these problems can sometimes be considerably simplified by a careful appeal to symmetry. Suppose we generate a new random process from  $(X, Y)$ :

$$(X', Y') = (-X, Y).$$

This new process is obtained from the old one simply by rotating  $A$  around the  $y$ -axis by  $180^\circ$ . However, it seems reasonable to conjecture, by appealing to symmetry, that  $(X', Y')$  also conforms to the distribution (4.21). This is quite true, and a proof follows easily after recognizing that the areas of  $E$  and  $A$  remain unchanged after this rotation.

This means that any expected value is the same for both  $(X, Y)$  and  $(X', Y')$ . We have, for example,

$$E[X] = E[X'] = E[-X] = -E[X]$$

This implies  $E[X] = 0$ . That  $E[Y] = 0$  follows from essential the same argument. This gives covariance

$$\text{cov}[X, Y] = E[(X - \mu_X)(Y - \mu_Y)] = E[XY].$$

Applying the same logic to  $XY$  we have

$$E[XY] = E[X'Y'] = E[-XY] = -E[XY],$$

so that  $\text{cov}[X, Y] = E[XY] = 0$ .

However,  $X$  and  $Y$  are not independent. To see this, note that  $E_1 = \{X > 0.99\}$  and  $E_2 = \{Y > 0.99\}$  are mutually exclusive events (they cannot both happen, since  $0.99^2 + 0.99^2 > 1$ ). This means  $P(E_1 \cap E_2) = 0$ . On the other hand  $P(E_1) > 0$  and  $P(E_2) > 0$ , so that

$$0 = P(E_1 \cap E_2) \neq P(E_1)P(E_2) > 0.$$

Thus, although the covariance between  $X$  and  $Y$  is zero, they are not independent. ■

#### 4.8.2 Conditional Probability and Independence

Independence between two RVs can be expressed also by conditional probabilities. Suppose  $X_1 \perp X_2$ . Let  $E_1, E_2 \subset \mathbb{R}$  be two sets. Then

$$\begin{aligned} P(X_1 \in E_1 \mid X_2 \in E_2) &= \frac{P(X_1 \in E_1, X_2 \in E_2)}{P(X_2 \in E_2)} \\ &= \frac{P(X_1 \in E_1)P(X_2 \in E_2)}{P(X_2 \in E_2)} \\ &= P(X_1 \in E_1). \end{aligned}$$

Intuitively, that  $P(X_1 \in E_1 \mid X_2 \in E_2) = P(X_1 \in E_1)$  implies that knowledge of the outcome of  $X_2$  does not alter the probability of any outcome involving  $X_1$ , equivalently, that  $X_1$  and  $X_2$  are independent. We have the counterpart to Theorem 2.1.

**Theorem 4.10** Given RVs  $X_1, X_2$

$$P(X_1 \in E_1 \mid X_2 \in E_2) = P(X_1 \in E_1)$$

and

$$P(X_2 \in E_2 \mid X_1 \in E_1) = P(X_2 \in E_2)$$

for all subsets  $E_1, E_2$  if and only if  $X_1 \perp X_2$ . ■

#### 4.9 Sums and Averages of Random Variables

Suppose  $X_1, X_2, \dots, X_n$  is a collection of random variables, not necessarily independent. There is often interest in the sum

$$S = \sum_{i=1}^n X_i,$$

and the *sample average*

$$\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n} = \frac{S}{n}.$$

We introduce the following shorthand,

$$\begin{aligned}\mu_i &= E[X_i], \quad i = 1, \dots, n, \\ \sigma_i^2 &= \text{var}[X_i] = E[(X_i - \mu_i)^2], \quad i = 1, \dots, n, \\ \sigma_{ij} &= \text{cov}[X_i, X_j] \\ &= E[(X_i - \mu_i)(X_j - \mu_j)], \quad i = 1, \dots, n, \quad j = 1, \dots, n.\end{aligned}\quad (4.22)$$

From Theorem 4.9 (iv) we have  $\text{cov}[X_i, X_i] = \text{var}[X_i]$  so that

$$\sigma_{ii} = \sigma_i^2.$$

From Theorem 4.9 (v) we have  $\text{cov}[X_i, X_j] = \text{cov}[X_j, X_i]$  so that

$$\sigma_{ij} = \sigma_{ji}.$$

#### 4.9.1 Expected Values of Sums and Averages

One convenient feature of expected values is that the expected value of a sum equals the sum of the expected values.

**Theorem 4.11** For any set of random variables  $X_1, \dots, X_n$  we have

$$E\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n E[X_i], \quad (4.23)$$

assuming all expected values are finite. ■

In order for Theorem 4.11 to hold, we must assume all expected values are finite. Suppose, to take a counterexample, we have a RV  $X$  for which  $E[X] = \infty$ . For example, if  $Y$  is any random variable for which  $P(Y = 0) > 0$ , then if  $X = 1/Y$ , we would have  $E[X] = \infty$ . However, we cannot conclude that

$$0 = E[X - X] = E[X] + E[-X] = \infty - \infty,$$

since  $\infty - \infty$  is not defined.

Also note that Theorem 4.11 holds even when the random variables are not independent. We similarly have:

**Theorem 4.12** For any set of random variables  $X_1, \dots, X_n$  we have

$$E\left[\bar{X}_n\right] = n^{-1} \sum_{i=1}^n E[X_i],$$

and if all RVs possess the same mean  $E[X_i] = \mu$  then

$$E\left[\bar{X}_n\right] = \mu. \quad \blacksquare$$

**Example 4.15** At a party,  $N$  men bring identical hats. At the end of the party each man brings home one of the hats chosen at random. Let  $S_N$  be the number of men who bring home their own hats. What is  $E[S_N]$ ?

Let  $X_i = 1$  if the  $i$ th man brings home his own hat, and let  $X_i = 0$  otherwise. This event has probability  $1/N$ , so that each  $X_i$  has PMF

$$p(0) = 1 - 1/N \text{ and } p(1) = 1/N$$

and from Definition 4.5 we have expected value

$$E[X_i] = 0 \times p(0) + 1 \times p(1) = 1/N.$$

Then,

$$E[S_N] = E[X_1] + \dots + E[X_N] = 1/N + \dots + 1/N = 1.$$

Interestingly, the answer does not actually depend on  $N$ . ■

#### 4.9.2 Variances of Sums and Averages

There will also be interest in the variance of a sum.

**Theorem 4.13** Suppose  $X_1, X_2, \dots, X_n$  is a collection of random variables, with sum

$$S = \sum_{i=1}^n X_i.$$

Using the notation of (4.22) the variance of the sum is given by

$$\text{var}[S] = \sum_{i=1}^n \sigma_i^2 + 2 \sum_{i < j} \sigma_{ij}.$$

If the RVs are independent, then the variance of the sum is given by

$$\text{var}[S] = \sum_{i=1}^n \sigma_i^2.$$

Finally, we always have

$$\text{var}[\bar{X}_n] = n^{-2} \text{var}[S],$$

and if the RVs are *iid* with variance  $\sigma^2$ , then

$$\text{var}[\bar{X}_n] = n^{-1} \sigma^2.$$

■

*Proof.* First note that

$$E[S] = \sum_{i=1}^n \mu_i.$$

Then

$$\begin{aligned}
\text{var}[S] &= E[(S - E[S])^2] \\
&= E\left[\left(\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i\right)^2\right] \\
&= E\left[\left(\sum_{i=1}^n (X_i - \mu_i)\right)^2\right] \\
&= E\left[\sum_{i=1}^n \sum_{j=1}^n (X_i - \mu_i)(X_j - \mu_j)\right] \\
&= \sum_{i=1}^n \sum_{j=1}^n E[(X_i - \mu_i)(X_j - \mu_j)] \\
&= \sum_{i=1}^n \sum_{j=1}^n \sigma_{ij} \\
&= \sum_{i=1}^n \sigma_i^2 + 2 \sum_{i < j} \sigma_{ij}, \tag{4.24}
\end{aligned}$$

which verifies (4.13). Note that we have made use of Equation 4.23 applied to random variables of the form  $(X_i - \mu_i)(X_j - \mu_j)$  which possess expected value

$$E[(X_i - \mu_i)(X_j - \mu_j)] = \sigma_{ij}.$$

We have also made use of Theorem 4.9 (v).

Next, suppose that the RVs are independent. Equation (4.13) follows from (4.13) and the fact that the covariances  $\sigma_{ij} = 0$  for  $i \neq j$  by independence and Theorem 4.9 (ii).

The rest of the proof follows directly from Theorem 4.7, noting that  $\bar{X}_n = n^{-1}S$ .  $\square$

## 4.10 Some Common Discrete Random Variables

Despite that fact that the definition of a random variable is quite general, the range of distributions encountered in most applications in probability and statistics is quite limited. This is because most random processes encountered in practice can be well approximated by a relatively small number of models.

### 4.10.1 Bernoulli Distribution

We say that  $U$  is a Bernoulli RV with parameter  $p \in [0, 1]$  if  $P(U = 1) = 1 - P(U = 0) = p$ . This definition suffices to restrict the outcomes of  $U$  to 0 and 1. Collections of Bernoulli RVs are often used to construct more elaborate random processes.

The expected value of  $U$  is easy to evaluate:

$$E[U] = 0 \times P(U = 0) + 1 \times P(U = 1) = p. \quad (4.25)$$

Since  $U$  must be 0 or 1, we have  $U^2 = U$ . The variance is then

$$\text{var}[U] = E[U^2] - E[U]^2 = E[U] - E[U]^2 = p - p^2 = p(1 - p). \quad (4.26)$$

We have already seen examples of Bernoulli RNs in Example 4.15.

We denote a Bernoulli RV  $U \sim \text{Bern}(p)$ .

#### 4.10.2 Binomial Distribution

One of the most important random variables is the *binomial random variable*. We may define a process of *binomial trials* according to the following definition:

1. An experiment with two outcomes (say, success or failure) is performed  $n$  times.
2. Each experiment has the same probability  $p$  of ending in a success.
3. Each experiment is independent of the other.

If we have such a process, let  $X$  be the total number of successes. Then  $X$  is a binomial random variable with parameters  $(n, p)$ . This is sometimes written  $X \sim \text{bin}(n, p)$ . We have  $S = \{0, 1, \dots, n\}$ . It can be shown that the distribution of  $X$  is given by the formula

$$P(X = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} = \binom{n}{k} p^k (1-p)^{n-k}$$

for any  $k = 0, 1, \dots, n$ . Many processes can be adequately modeled using binomial trials. For example, the number of boys in a family, the number of defective products detected by an inspector, the number of winners at a roulette wheel, and so on.

It is useful to recognize that  $X \sim \text{bin}(n, p)$  is the sum of  $n$  independent Bernoulli RVs  $U_1, \dots, U_n$  with mean  $p$ . Recalling equation (4.23) from Section 4.9.1 we can conclude that

$$E[X] = \sum_{i=1}^n E[U_i] = np.$$

The variance follows from Theorem 4.13 and the variance of the Bernoulli RV given in Equation 4.26

$$\text{var}[X] = \sum_{i=1}^n \text{var}[U_i] = np(1-p).$$

**Example 4.16** Using past experience, a salesperson knows that a customer who enters a certain store has a probability of 0.2 of making a purchase. Suppose there are 5 customers in this store at the moment. What is the probability that at least 2 purchases will be made?

We will first calculate the probability that no more than 1 purchase is made. Since this is the complement of the event in question, we can obtain our final answer by subtracting this probability from 1. We then have

$$P(X \leq 1) = P(X = 0) + P(X = 1).$$

Using the binomial formula with  $p = 0.2$  and  $n = 5$  we get

$$\begin{aligned} P(X = 0) &= \frac{5!}{0!5!} 0.2^0 0.8^5 \\ &= 0.8^5 \\ &= 0.328 \end{aligned}$$

and

$$\begin{aligned} P(X = 1) &= \frac{5!}{1!4!} 0.2^1 0.8^4 \\ &= 5 \times .2^1 \times 0.8^4 \\ &= 0.410 \end{aligned}$$

so that

$$\begin{aligned} P(X \leq 1) &= 0.328 + 0.410 \\ &= 0.738. \end{aligned}$$

This means that

$$\begin{aligned} P(X \geq 2) &= 1 - P(X \leq 1) \\ &= 1 - 0.738 \\ &= 0.262. \end{aligned}$$

The probability that at least 2 of the 5 customers makes a purchase is therefore 0.262. ■

**Example 4.17** Suppose we have an opinion poll of 1000 randomly selected respondents, of which 490 are male (that is, 49%). We suspect that males are under-represented, that is, the number of males should be closer to 500. Which of the following probabilities would we be interested in?

1. That there are exactly 490/1000 male respondents. ( $P \approx 0.021$ )
2. That there are exactly 490/1000 male or 490/1000 female respondents. ( $P \approx 0.042$ )

3. That there are no more than 490/1000 male respondents. ( $P \approx 0.274$ )
4. That there are no more than 490/1000 male or 490/1000 female respondents. ( $P \approx 0.548$ )

It helps to know that the probability that the sample is *exactly balanced* (that is, that there are exactly 500/1000 male and exactly 500/1000 female respondents is  $P \approx 0.025$ ). ■

### 4.10.3 Poisson Random Variables

When a random variable  $X$  represents a count of some event in time or some phenomenon over space, this random variable is often modelled using a *Poisson* random variable, which has distribution

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \dots$$

A Poisson RV is denoted  $X \sim \text{Pois}(\lambda)$ .

The expected value can be calculated by a change of variable strategy:

$$\begin{aligned} E[X] &= \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} \\ &= \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} e^{-\lambda} \\ &= \sum_{k=0}^{\infty} \frac{\lambda^{k+1}}{k!} e^{-\lambda} \\ &= \lambda \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} \\ &= \lambda, \end{aligned}$$

exploiting the fact that the sum of the PMF reappears in the second last line above. To calculate the variance, we first calculate the related expectation:

$$\begin{aligned} E[X(X - 1)] &= \sum_{k=0}^{\infty} k(k-1) \frac{\lambda^k}{k!} e^{-\lambda} \\ &= \sum_{k=2}^{\infty} \frac{\lambda^k}{(k-2)!} e^{-\lambda} \\ &= \sum_{k=0}^{\infty} \frac{\lambda^{k+2}}{k!} e^{-\lambda} \\ &= \lambda^2 \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} \\ &= \lambda^2. \end{aligned}$$

The second moment is then readily obtained:

$$E[X^2] = E[X^2] - E[X] + E[X] = E[X(X - 1)] + E[X] = \lambda^2 + \lambda$$

along with the variance

$$\text{var}[X] = E[X^2] - E[X]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

Typically, the number of arrivals at a bank during a fixed time interval will have approximately a Poisson distribution. Similarly the number of flaws in a sheet of plywood of fixed area may also have a Poisson distribution.

**Example 4.18** Customers arrive at a bank at a rate  $\lambda$  customers per minute. If certain assumptions are met, the number of customers  $X$  arriving during a time interval of  $T$  minutes will have a Poisson distribution with expected value  $E[X] = \lambda T$ . That is, an average of  $\lambda T$  customers will arrive in a time interval of length  $T$ , with distribution function

$$P(X = k) = \frac{(\lambda T)^k}{k!} e^{-\lambda T}, \quad k = 0, 1, 2, \dots$$

This is known as a *Poisson process*. ■

If  $X \sim \text{bin}(n, p)$ , with  $p$  very small, and  $n$  very large, then  $X$  will also have approximately a Poisson distribution with average value  $\lambda = np$ .

**Example 4.19** What is the probability that there will be at least 3 birthdays among 100 people?

Let  $X$  be the number of birthdays. Then  $X \sim \text{bin}(100, 1/365)$ , so that

$$\begin{aligned} P(X \geq 3) &= 1 - P(X = 0) - P(X = 1) - P(X = 2) \\ &= 1 - \binom{100}{0} (1/365)^0 (364/365)^{100} \\ &\quad - \binom{100}{1} (1/365)^1 (364/365)^{99} \\ &\quad - \binom{100}{2} (1/365)^2 (364/365)^{98} \\ &= 0.002727. \end{aligned}$$

On the other hand, if we use the Poisson approximation we set

$$\lambda = 100/365$$

giving

$$\begin{aligned} P(X \geq 3) &= 1 - P(X = 0) - P(X = 1) - P(X = 2) \\ &\approx 1 - \frac{(100/365)^0}{0!} e^{-100/365} \\ &\quad - \frac{(100/365)^1}{1!} e^{-100/365} \\ &\quad - \frac{(100/365)^2}{2!} e^{-100/365} \\ &= 0.002795 \end{aligned}$$

giving the same answer to 2 significant digits. ■

#### 4.10.4 Geometric Random Variable

The random variable shown in Example 4.9 is a *geometric random variable*. Suppose  $U_1, U_2, \dots$  is an infinite sequence of independent Bernoulli random variables with parameter  $p$ . Let  $X$  be the first index  $i$  for which  $U_i = 1$ . We interpret  $X$  as being the number of Bernoulli trials required to observe a ‘success’ (or, a ‘1’). Then the PMF is given by the following argument (review Example 4.9):

$$\begin{aligned} P(X = k) &= P(U_1 = 0, \dots, U_{k-1} = 0, U_k = 1) \\ &= P(U_1 = 0) \times \dots \times P(U_{k-1} = 0) \times P(U_k = 1) \\ &= (1 - p)^{k-1} p, \quad k = 1, 2, \dots \end{aligned}$$

Note that in some texts the geometric random variable is defined as the number of 0s appearing before the first 1, which would be  $X - 1$  under the definition used here.

We denote a geometric RV  $X \sim \text{geom}(p)$ .

The moments may be evaluated using the geometric series (Section B.5):

$$\begin{aligned} E[X] &= \sum_{i=1}^{\infty} ip(1 - p)^{i-1} = \frac{p}{(1 - (1 - p))^2} = \frac{1}{p}, \\ E[(X + 1)X] &= \sum_{i=1}^{\infty} (i + 1)ip(1 - p)^{i-1} = \frac{2p}{(1 - (1 - p))^3} = \frac{2}{p^2}. \end{aligned}$$

This leads to variance

$$\begin{aligned} \text{var}[X] &= E[X^2] - E[X]^2 \\ &= E[(X + 1)X] - E[X] + E[X]^2 \\ &= \frac{2}{p^2} - \frac{1}{p} + \frac{1}{p^2} \\ &= \frac{1 - p}{p^2}. \end{aligned}$$

The CDF of the geometric distribution is conveniently determined by the tail probability

$$\begin{aligned} \bar{F}_X(k) &= P(X > k) \\ &= \sum_{i=k+1}^{\infty} p(1 - p)^{i-1} \\ &= p(1 - p)^k + p(1 - p)^{k+1} + \dots \\ &= p(1 - p)^k (1 + (1 - p) + (1 - p)^2 + \dots) \\ &= (1 - p)^k, \end{aligned}$$

from the geometric series. So the CDF is

$$F_X(k) = 1 - (1 - p)^k. \quad (4.27)$$

**Example 4.20** The geometric random variable possesses the interesting *memoryless property*. Suppose you have been repeatedly playing a game of chance with a probability  $p$  of winning. After  $k$  games you have not yet won. We assume the outcomes are independent. Is the amount of time until you win, starting from that point, different from the time to win when you started to play? Many are tempted to believe that prior losses shorten the expected time to future wins, as though the number of losses is somehow fixed. The question is answered by the following conditional probability:

$$P(X > k + t \mid X > k) = \frac{P(X > k + t)}{P(X > k)} = \frac{(1 - p)^{k+t}}{(1 - p)^k} = (1 - p)^t, \quad (4.28)$$

using CDF (4.27). This is the probability that at least  $t$  further losses precede a win following  $k$  consecutive losses. It is exactly the probability that at least  $t$  losses precede a win at the beginning of play. In other words, the waiting time for a win following  $k$  losses does not depend on  $k$  in any way. Equation (4.28) is a statement of the memoryless property. ■

#### 4.10.5 Negative Binomial Random Variable

The geometric random variable can be generalized in the following way. Suppose we are given the same Bernoulli trials used to define the geometric RV in Section 4.10.4, but we then let  $X$  be the number of Bernoulli trials required to observe  $r$  1's. Suppose  $X = k$ . This means that in the first  $k$  trials there are  $r$  1's and  $k-r$  0's. However, the final element of the sequence must be a 1. Otherwise, the 0's and 1's in the first  $k-1$  elements can be in any order. Using the enumeration principles discussed in Section , we can see that there must be

$$N = \binom{k-1}{r-1}$$

such sequences. In addition, any such sequence, consisting of  $r$  1's and  $k-r$  0's must have probability  $p^r(1-p)^{k-r}$ . We conclude that  $X$  has PMF

$$p_X(k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}.$$

We denote a negative binomial RV  $X \sim nb(r, p)$ .

To calculate the mean and variance, we simply note that by construction  $X$  is a sum of  $r$  independent geometric random variables of mean  $1/p$  and variance  $(1-p)/p^2$ . Using Equation (4.23) and Theorem 4.13 we have

$$\begin{aligned} E[X] &= \frac{r}{p} \\ \text{var}[X] &= \frac{r(1-p)}{p^2} \end{aligned}$$

## 4.11 Some Common Continuous Random Variables

As in the case of discrete RVs there is a relatively small set of continuous RVs which appear naturally in many important probability models.

### 4.11.1 The Uniform Distribution

We say that  $U$  has a *uniform distribution* on interval  $[a, b]$  if the density function is constant on  $[a, b]$  and zero otherwise. Since the density must integrate to one on  $\mathbb{R}$  we have for some constant  $c$

$$1 = \int f_U(u)du = \int_a^b cdu = c(b-a).$$

This means  $c = (b-a)^{-1}$  giving the density precisely as

$$f_U(u) = \begin{cases} (b-a)^{-1} & ; \quad u \in [a, b] \\ 0 & ; \quad u \notin [a, b] \end{cases} \quad (4.29)$$

We denote a uniform RV  $U \in \text{unif}[a, b]$ .

The moments are given by the integral

$$E[U^k] = \int_a^b u^k / (b-a) du = \frac{u^{k+1}}{(k+1)(b-a)} \Big|_a^b = \frac{b^{k+1} - a^{k+1}}{(k+1)(b-a)}.$$

In particular, for  $k = 1, 2$  we have

$$\begin{aligned} E[U] &= \frac{b^2 - a^2}{2(b-a)} = \frac{(b-a)(b+a)}{2(b-a)} = \frac{a+b}{2} \\ E[U^2] &= \frac{b^3 - a^3}{3(b-a)} = \frac{(b-a)(a^2 + ab + b^2)}{3(b-a)} = \frac{a^2 + ab + b^2}{3}. \end{aligned}$$

The mean of  $U$  is, as would be expected by symmetry, the midpoint of the interval  $[a, b]$ . The variance is then

$$\text{var}[U] = E[U^2] - E[U]^2 = \frac{a^2 + ab + b^2}{3} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12}.$$

The CDF follows directly from the careful integration of the piecewise constant density (4.29)

$$F_U(u) = \begin{cases} 0 & ; \quad u \leq a \\ \frac{u-a}{b-a} & ; \quad a < u \leq b \\ 1 & ; \quad u > b \end{cases}$$

### 4.11.2 The Exponential Distribution

We say  $X$  possesses an *exponential distribution* if for some  $\lambda > 0$  it has density

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & ; \quad x \geq 0 \\ 0 & ; \quad x < 0 \end{cases}.$$

Then  $X$  is a *positive random variable*, that is  $P(X \geq 0) = 1$ . We denote the exponential RV  $X \sim \exp(\lambda)$ .

The mean and variance may be evaluated using the gamma function  $\Gamma(t)$  defined in Section B.7. Using Equation (B.4) (setting  $t = 2$ ,  $a = \lambda$ ) we have

$$\begin{aligned} E[X] &= \int_0^\infty x \lambda e^{-\lambda x} dx \\ &= \lambda \int_0^\infty x \lambda e^{-\lambda x} dx \\ &= \lambda \frac{1}{\lambda^2} \Gamma(2) \\ &= \lambda^{-1}, \end{aligned}$$

since  $\Gamma(2) = 1! = 1$ . The second moment is (setting  $t = 3$ ,  $a = \lambda$ )

$$\begin{aligned} E[X^2] &= \int_0^\infty x^2 \lambda e^{-\lambda x} dx \\ &= \lambda \int_0^\infty x^2 \lambda e^{-\lambda x} dx \\ &= \lambda \frac{1}{\lambda^3} \Gamma(3) \\ &= 2\lambda^{-2}, \end{aligned}$$

since  $\Gamma(3) = 2! = 2$ . The variance follows directly,

$$\text{var}[X] = E[X^2] - E[X]^2 = 2\lambda^2 - \lambda^2 = 1/\lambda^2.$$

At this point, it is worth noting the resemblance of the exponential density to the geometric density (Section 4.10.4). First, the CDF is naturally calculated following the tail probability, for  $t > 0$

$$\bar{F}_X(t) = P(X > t) = \int_t^\infty \lambda e^{-\lambda x} = -e^{-\lambda x} \Big|_t^\infty = e^{-\lambda t},$$

so that the CDF is

$$F_X(t) = \begin{cases} 0 & ; \quad t < 0 \\ 1 - e^{-\lambda t} & ; \quad t \geq 0 \end{cases}. \quad (4.30)$$

Second, the exponential RV possesses the same type of memoryless property demonstrated in Example 4.20.

**Theorem 4.14** An exponentially distributed RV  $X \sim \exp(\lambda)$  possesses the memoryless property

$$P(X > t + s \mid X > s) = P(X > t), \quad (4.31)$$

for any positive values  $s, t$ . ■

*Proof.* . We have directly from (4.30)

$$\begin{aligned} P(X > t + s \mid X > s) &= \frac{P(X > t + s, X > s)}{P(X > s)} \\ &= \frac{P(X > t + s)}{P(X > s)} \\ &= \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} \\ &= e^{-\lambda t} \end{aligned}$$

from which (4.31) follows. □

## 4.12 The Normal Distribution

Perhaps the most important distribution in statistics is the *normal distribution*. The familiar “bell shape” of its density is shown in (Figure 4.7).

In order to completely define a normal random variable we need to specify the mean  $\mu$  and the variance  $\sigma^2$ . We then write

$$X \sim N(\mu, \sigma^2).$$

The support of  $X$  is  $S_X = (-\infty, \infty)$ . The density function of  $X$  is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}, \quad x \in (-\infty, \infty)$$

and so the CDF is

$$F_X(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(u-\mu)^2/(2\sigma^2)} du, \quad x \in (-\infty, \infty). \quad (4.32)$$

It can be verified, using techniques from calculus, that  $E[X] = \mu$  and  $\text{var}[X] = \sigma^2$ .

One technical matter is worth noting. The value of the *definite integral*

$$1 = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(u-\mu)^2/(2\sigma^2)} du$$

can be verified analytically. Unfortunately, the *indefinite integral* used in Equation (4.32) to define the CDF does not have closed form solution (that is, there is no convenient formula with which to calculate  $F_X(x)$ ). This turns out to be true for a number of important distributions used in statistics. We will see in the next section how the CDF can be calculated.

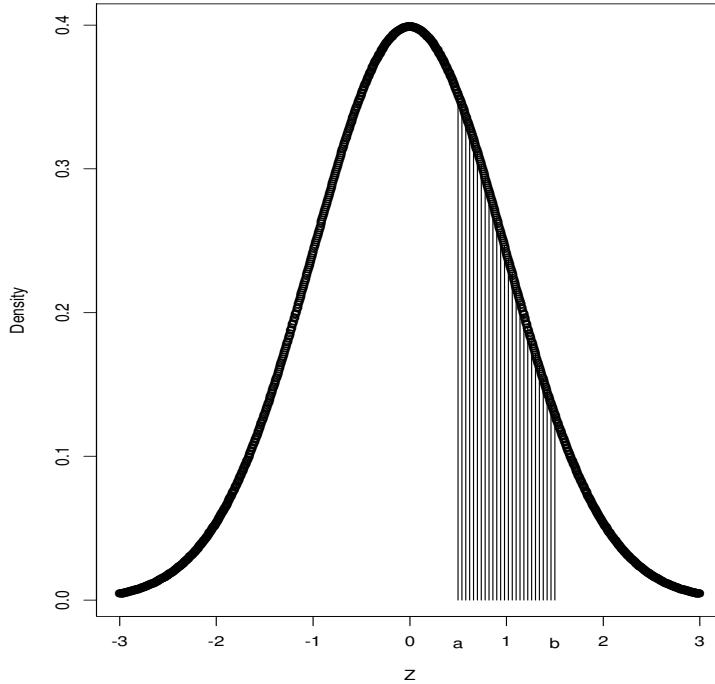


Figure 4.7: Normal density. Shaded area is equal to  $P(a \leq Z \leq b) = \int_a^b f_Z(z) dz$

#### 4.12.1 Calculating Normal Probabilities

An important special case of the normal distribution is given by

$$Z \sim N(0, 1),$$

with density function

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad z \in (-\infty, \infty).$$

This is referred to as the *standard normal* or *unit normal* distribution. Since no closed form expression exists for the CDF of  $Z$  we used standard normal tables. These tables give values for  $P(Z \leq z)$ , the shaded area in Figure 4.8.

Tables A.1 and A.2 (Appendix A) give  $P(Z \leq z)$  for values of  $z$  in between -3.49 and 3.49. We first round  $z$  to two decimal places, using 3 digits in all. We locate the first two digits along the left column, and then we locate the final digit along the top. The probability is located at the intersection of the selected row and column.

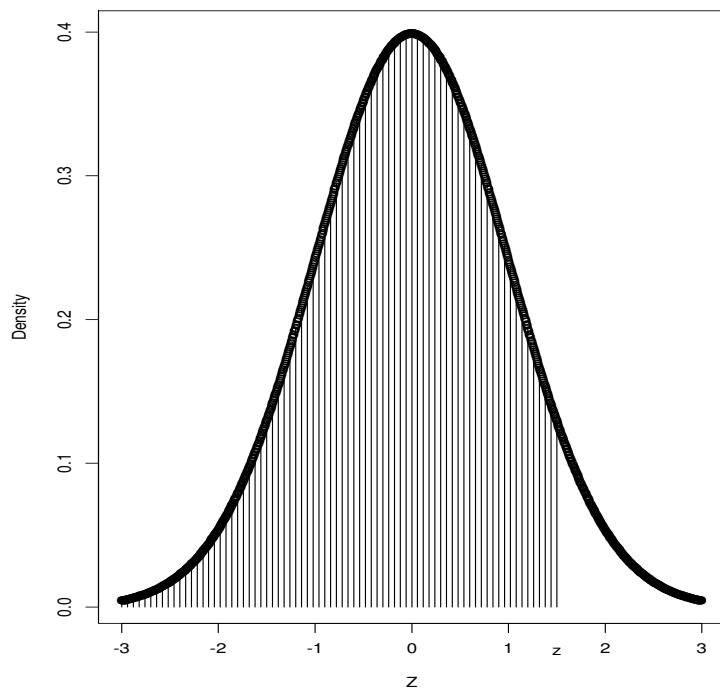


Figure 4.8: Standard normal distribution probabilities. Shaded area is equal to  $P(X \leq z) = F_Z(z)$

**Example 4.21** If  $Z \sim N(0, 1)$  then to calculate  $P(Z \leq -0.49)$  we first locate -0.4 along the left column. We then locate 0.09 along the top. The value at the intersection is then 0.3121. ■

The CDF can be used to calculate probabilities of various forms, using the rules of elementary probability to get, for example,

$$\begin{aligned} P(Z > z) &= 1 - P(Z \leq z) = 1 - F_Z(z), \text{ or} \\ P(a < Z \leq b) &= P(Z \leq b) - P(Z \leq a) = F_Z(b) - F_Z(a). \end{aligned}$$

These rules allow us to use the tables to calculate any probability involving  $Z$ .

Note that  $P(Z = z) = 0$  for any  $z$ . That is, the probability that  $Z$  equals exactly  $z$  is 0, at least theoretically. This means that we may write

$$\begin{aligned} P(Z \leq z) &= P(Z < z) \\ P(a < Z \leq b) &= P(a \leq Z < b) \\ &= P(a \leq Z \leq b) \\ &= P(a < Z < b) \end{aligned}$$

and so on.

It is sometimes convenient to rely on the symmetry about 0 of the standard normal density. For example, if  $c > 0$  then

$$P(Z \leq -c) = P(Z \geq c)$$

This means

$$\begin{aligned} P(|Z| \leq c) &= P(-c \leq Z \leq c) \\ &= 1 - P(Z < -c) - P(Z > c) \\ &= 1 - P(Z \leq -c) - P(Z \geq c) \\ &= 1 - 2P(Z \leq -c) \\ &= 1 - 2F_Z(-c). \end{aligned}$$

#### 4.12.2 Linear Transformations of Normal RVs

One important property of normal RVs is that their linear transformations are also normally distributed. This being the case, from Theorems 4.5 and 4.7 we have

$$X \sim N(\mu, \sigma^2) \text{ implies } Y = aX + b \sim N(a\mu + b, a^2\sigma^2).$$

In particular, if  $Z \sim N(0, 1)$  then  $X \sim N(\mu, \sigma^2)$  can be constructed as

$$X = \sigma Z + \mu.$$

### 4.12.3 Calculating General Normal Probabilities

We need to be able to calculate the probability of any normal random variable. In order to do so we rely on a simple rule. If

$$X \sim N(\mu, \sigma^2)$$

and

$$Z = \frac{X - \mu}{\sigma}$$

then

$$Z \sim N(0, 1).$$

The CDF of  $X \sim N(\mu, \sigma^2)$  is then

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= P\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) \\ &= P\left(Z \leq \frac{x - \mu}{\sigma}\right) \\ &= F_Z\left(\frac{x - \mu}{\sigma}\right). \end{aligned}$$

In this way, the CDF  $F_X(x)$  of any normal random variable can be evaluated by first calculating  $z = (x - \mu)/\sigma$ , then evaluating  $F_Z(z)$ , using tables or software. The result of this type of transformation applied to a random variable  $X$  or quantile  $x$  is sometimes referred to as a *z-score*  $Z = (X - \mu)/\sigma$  or  $z = (x - \mu)/\sigma$ .

We see how this is used in the following example.

**Example 4.22** Suppose that the scores for a college entrance test are normally distributed with mean 600 and standard deviation 100. What proportion of students are between 650 and 700?

We first note that

$$X \sim N(600, 100^2)$$

so that

$$\begin{aligned} \mu &= 600 \\ \sigma &= 100. \end{aligned}$$

We can then write

$$\begin{aligned} P(650 < X < 700) &= P\left(\frac{650 - \mu}{\sigma} < \frac{X - \mu}{\sigma} < \frac{700 - \mu}{\sigma}\right) \\ &= P\left(\frac{650 - 600}{100} < Z < \frac{700 - 600}{100}\right) \\ &= P(0.5 < Z < 1.0) \end{aligned}$$

where we set

$$Z = \frac{X - \mu}{\sigma},$$

and note that  $Z \sim N(0, 1)$ , so we may use the standard normal tables. We then have

$$\begin{aligned} P(0.5 < Z < 1.0) &= P(Z < 1.0) - P(Z < 0.5) \\ &= 0.8413 - 0.6915 \\ &= 0.1498 \end{aligned}$$

so that 14.98% of test scores are between 650 and 700. ■

#### 4.12.4 Normal Quantiles

We can also use the table to calculate quantiles (see Definition 4.4). If we wish to calculate the  $\alpha$  quantile, we look up  $\alpha$  in the table, and find the value of  $z$  to which it corresponds. Then  $z$  is the desired quantile. As a matter of notation we denote this value  $Z_\alpha$ . In principle we can calculate the quantile for any normal distribution, but we may directly use the tables only for standard normal quantiles. Fortunately, we can relate the quantiles of the standard normal distribution to the quantiles of an arbitrary  $N(\mu, \sigma^2)$  distribution through the formulae

$$\begin{aligned} Z_\alpha &= \frac{X_\alpha - \mu}{\sigma} \\ X_\alpha &= \mu + \sigma Z_\alpha \end{aligned} \tag{4.33}$$

where  $X_\alpha$  is the  $\alpha$ -quantile of a  $N(\mu, \sigma^2)$  distribution.

To summarize, a standard quantile is obtained by solving

$$\alpha = F_Z(Z_\alpha),$$

after which any normal quantile can be obtained by the formula (4.33).

**Example 4.23** In the previous example, what score is 95th percentile? From the tables we find that

$$P(Z \leq 1.65) = 0.9505$$

and

$$P(Z \leq 1.64) = 0.9495$$

which means that

$$Z_{0.9505} = 1.65$$

and

$$Z_{0.9495} = 1.64.$$

Note that we cannot locate  $Z_{0.95}$  directly from the table, but we can assume that it is between  $Z_{0.9495}$  and  $Z_{0.9505}$ . Therefore, as an approximation we take

$$\begin{aligned} Z_{0.95} &\approx \frac{Z_{0.9495} + Z_{0.9505}}{2} \\ &= \frac{1.64 + 1.65}{2} \\ &= 1.645. \end{aligned}$$

This type of procedure is known as *interpolation*. Then using the conversion formula gives

$$\begin{aligned} X_{0.95} &= \mu + \sigma Z_{0.95} \\ &\approx 600 + 100 \times 1.645 \\ &= 764.5 \end{aligned}$$

so that the 95th percentile score is 764.5. ■

#### 4.12.5 Critical Values of the Normal Distribution

Critical values for the normal distribution are commonly used in statistical procedures. This is the value  $z_\alpha$  above which the area underneath the standard normal curve is equal to  $\alpha$  (Section 4.4.1). The critical value is related to the quantile through the relationship

$$z_\alpha = Z_{1-\alpha}.$$

This is illustrated in Figure 4.9.

### 4.13 Central Limit Theorem

Suppose we have an *iid* sample  $X_1, \dots, X_n$  (Definition 4.9). Then the *law of large numbers* states that there is some number  $\mu$  such that

$$\bar{X}_n \approx \mu,$$

and that the approximation becomes more accurate as  $n$  increases. In this case  $\mu$  is the mean of the distribution.

One of the most important theorems in statistics is the *central limit theorem*. It states that if we have a list of *iid* RVs  $X_1, \dots, X_n$ , with mean  $\mu$  and variance  $\sigma^2$ , then approximately

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

with the approximation becoming more accurate as  $n$  increases.

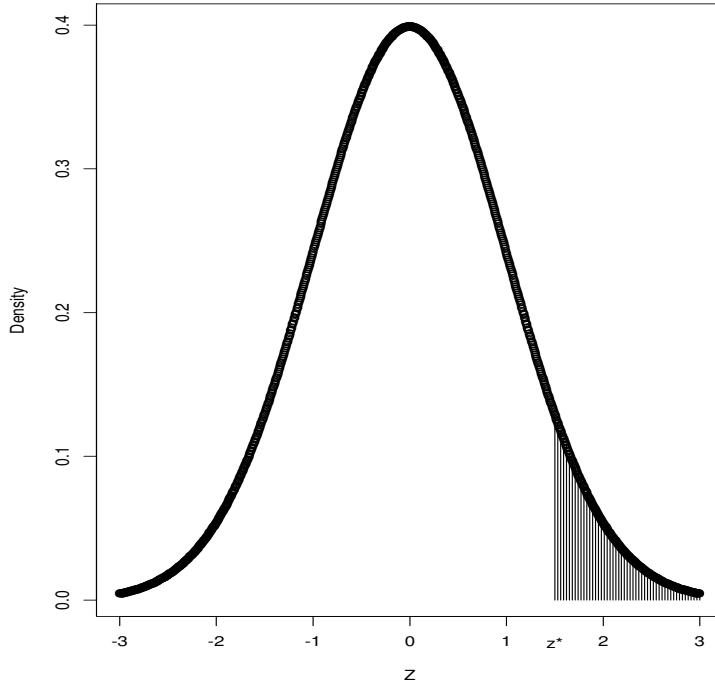


Figure 4.9: Critical values for a normal distribution. Shaded area is equal to  $\alpha = P(X \geq z_\alpha) = 1 - F_Z(z_\alpha)$ .

It is important to note that this result holds whether or not the individual random variables  $X_i$  are themselves normally distributed. This allows considerable simplification when dealing with sample averages. Also, the statement is exact when they are normally distributed.

We finally note that the central limit theorem can take alternative forms. Sometimes, it is more natural to consider the sum

$$Y = \sum_{i=1}^n X_i.$$

Under the same assumptions we would have, approximately,

$$Y \sim N(n\mu, n\sigma^2),$$

since  $E[Y] = n\mu$  and  $\text{var}[Y] = n\sigma^2$  (Theorems 4.5 and 4.7). In general, when we say  $Y$  is approximately normal, we mean  $Y \sim N(E[Y], \text{var}[Y])$ , approximately.

**Example 4.24** Suppose the true mean adult weight of a species of fish is 1.2 pounds, and that the standard deviation of adult weights is 0.5 pounds.

What is the probability that the average weight of a sample of 90 adult fish will be within 0.1 pound of the true mean?

The average weight is the sample mean  $\bar{X}_{90}$ . By the central limit theorem we have, approximately,

$$\begin{aligned}\bar{X}_{90} &\sim N\left(\mu, \frac{\sigma^2}{n}\right) \\ &\sim N\left(1.2, \frac{0.5^2}{90}\right)\end{aligned}$$

so that the standard deviation of  $\bar{X}_{90}$  is

$$\begin{aligned}\sigma_{\bar{X}_{90}} &= \frac{\sigma}{\sqrt{n}} \\ &= \frac{0.5}{\sqrt{90}} \\ &= 0.053.\end{aligned}$$

The desired probability is then

$$\begin{aligned}P(1.2 - 0.1 < \bar{X}_{90} < 1.2 + 0.1) &= P(1.1 < \bar{X}_{90} < 1.3) \\ &= P\left(\frac{1.1 - \mu}{\sigma_{\bar{X}_{90}}} < \frac{\bar{X}_{90} - \mu}{\sigma_{\bar{X}_{90}}} < \frac{1.3 - \mu}{\sigma_{\bar{X}_{90}}}\right) \\ &= P\left(\frac{1.1 - 1.2}{0.053} < Z < \frac{1.3 - 1.2}{0.053}\right) \\ &= P(-1.89 < Z < 1.89) \\ &= 1 - 2F_Z(-1.89) \\ &= 1 - 2 \times 0.0294 \\ &= 0.9412\end{aligned}$$

where  $Z \sim N(0, 1)$ . Thus, the probability that the sample average is within 0.1 pound of the true mean is 0.9412. ■

## 4.14 Normal Approximation to the Binomial

Recall the binomial random variable

$$X \sim \text{bin}(n, p).$$

When  $n$  is "large enough" then  $X$  will have approximately a normal distribution with mean and standard deviation

$$\begin{aligned}\mu_X &= np \\ \sigma_X &= \sqrt{np(1-p)}.\end{aligned}$$

We expect this because  $X$  is the sum of  $n$  independent random variables (with a  $\text{bin}(1, p)$  distribution).

Alternatively, we may write for the  $z$ -score

$$Z = \frac{X - np}{\sqrt{np(1-p)}} \sim N(0, 1), \text{ approximately.}$$

**Example 4.25** To test a dice for fairness it is tossed 600 times. On average, if the dice is fair, we would roll 6 100 times. Suppose we actually get 112 6's. Can this be expected if the dice is fair?

We need to calculate  $P(X \geq 112)$ . If this number is very small, then we conclude that the coin is not fair. We have

$$X \sim \text{bin}(600, 1/6),$$

with

$$\begin{aligned}\mu_X &= np = 100 \\ \sigma_X &= \sqrt{np(1-p)} = 3.73.\end{aligned}$$

Using the normal approximation, we get

$$\begin{aligned}P(X \geq 112) &= P\left(Z \geq \frac{112 - \mu_X}{\sigma_X}\right) \\ &= P\left(Z \geq \frac{112 - 100}{3.73}\right) \\ &= P(Z \geq 3.22) \\ &= 0.0006\end{aligned}$$

from which we conclude that 112 6's is unlikely with a fair dice. ■

An alternative way of expressing the normal approximation is by using the *proportion of successes*

$$\hat{p} = \frac{X}{n}$$

rather than  $X$  itself. This is essentially a sample mean.

In this case  $\hat{p}$  is also approximately normally distributed with

$$\begin{aligned}\mu_{\hat{p}} &= p \\ \sigma_{\hat{p}} &= \sqrt{\frac{p(1-p)}{n}},\end{aligned}$$

or alternatively

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1), \text{ approximately.}$$

**Example 4.26** Suppose the true support of a candidate running for office is given by proportion  $p = 0.46$ . What is the probability that an opinion poll using 200 respondents will estimate support at greater than 0.5?

Letting  $\hat{p}$  be the proportion of support in the poll, we may apply the normal approximation. The mean and standard deviation are given by

$$\begin{aligned}\mu_{\hat{p}} &= p = 0.46 \\ \sigma_{\hat{p}} &= \sqrt{\frac{p(1-p)}{n}} = 0.0352.\end{aligned}$$

We then calculate

$$\begin{aligned}P(\hat{p} \geq 0.5) &= P\left(Z \geq \frac{0.5 - \mu_{\hat{p}}}{\sigma_{\hat{p}}}\right) \\ &= P\left(Z \geq \frac{0.5 - 0.46}{0.0352}\right) \\ &= P(Z \geq 1.14) \\ &= 0.1271.\end{aligned}$$

We conclude that it is not unlikely that a true support of 46% may be mistaken for a majority support with a sample size of 200.

Suppose the size of the survey is increased to  $n = 1,000$ . The mean and standard deviation become

$$\begin{aligned}\mu_{\hat{p}} &= p = 0.46 \\ \sigma_{\hat{p}} &= \sqrt{\frac{p(1-p)}{n}} = 0.0158.\end{aligned}$$

The probability now becomes

$$\begin{aligned}P(\hat{p} \geq 0.5) &= P\left(Z \geq \frac{0.5 - \mu_{\hat{p}}}{\sigma_{\hat{p}}}\right) \\ &= P\left(Z \geq \frac{0.5 - 0.46}{0.0158}\right) \\ &= P(Z \geq 2.53) \\ &= 0.0057.\end{aligned}$$

We are much less likely to mistake a true support of 46% for a majority with a sample of size 1,000. ■

We stated previously that for the normal approximation to be applicable to a binomial random variable we required  $n$  to be "large enough". How large is "large enough"?

There are several rules of thumb available to check this. Generally, they also depend on the value of  $p$ . One of the most commonly used is the following:

If both

$$np \geq 5$$

and

$$n(1-p) \geq 5$$

hold, then the normal approximation will hold. (Some text books will use the number 10 in place of 5).

#### 4.14.1 Correction Rule for Normal Approximation to the Binomial

There are several ways to compare  $X_{bin} \sim bin(n, p)$  to  $X_{norm} \sim N(np, np(1-p))$ . We already have

$$E[X_{bin}] = E[X_{norm}] = np \text{ and } \text{var}[X_{bin}] = \text{var}[X_{norm}] = np(1-p),$$

so the important question is whether or not we have

$$P(X_{bin} \leq x) \approx P(X_{norm} \leq x).$$

In the previous section, we introduced a rule of thumb  $np \geq 5$  and  $n(1-p) \geq 5$ , under which the approximation will be reasonably good.

It is worth exploring the issue further. One of the problems in comparing  $X_{bin}$  to  $X_{norm}$  is that  $X_{bin}$  is discrete, and so does not possess a density. However, we can always represent the distribution of  $X_{bin}$  graphically as a histogram (which we will discuss in more detail in Section 9.2). Essentially, we force a discrete RV to become continuous by assigning to each outcome  $0, 1, 2, \dots, n$  a *class interval* of width 1, centered at that outcome. Then a rectangle is constructed using the class interval as the base, and with height equal to the probability of that outcome. The result is a true density function satisfying Definition 4.2, with probability  $P(X_{bin} = k)$  equal to the area under the histogram over the interval  $(k - 0.5, k + 0.5)$ . We can then compare the histogram directly to a normal density superimposed onto it. If the densities are close, then the approximation

$$P(X_{bin} = k) \approx P(k - 0.5 \leq X_{norm} \leq k + 0.5)$$

would be appropriate. This means the CDF of  $X_{bin}$  should use the approximation

$$F_{X_{bin}}(k) = P(X_{bin} \leq k) \approx P(X_{norm} \leq k + 0.5) = F_{X_{norm}}(k + 0.5) \quad (4.34)$$

rather than  $F_{X_{norm}}(k)$ . This can be seen in the following example.

**Example 4.27** Suppose  $X_{bin} \sim bin(n, p)$  for  $n = 5$  and  $p = 1/2$ . The rule of thumb of the previous section does not hold, since  $np = n(1-p) = 2.5 < 5$ . However, if we apply the correction given in Equation (15.1) we obtain a very close approximation, with  $P(X_{bin} \leq 1) \approx 0.187$  and  $P(X_{norm} \leq 1.5) \approx 0.186$ . In contrast,  $P(X_{norm} \leq 1) \approx 0.0899$ . In general, for small values of  $n$ , this correction procedure results in a considerable improvement in the normal approximation. See Figure 4.10 ■

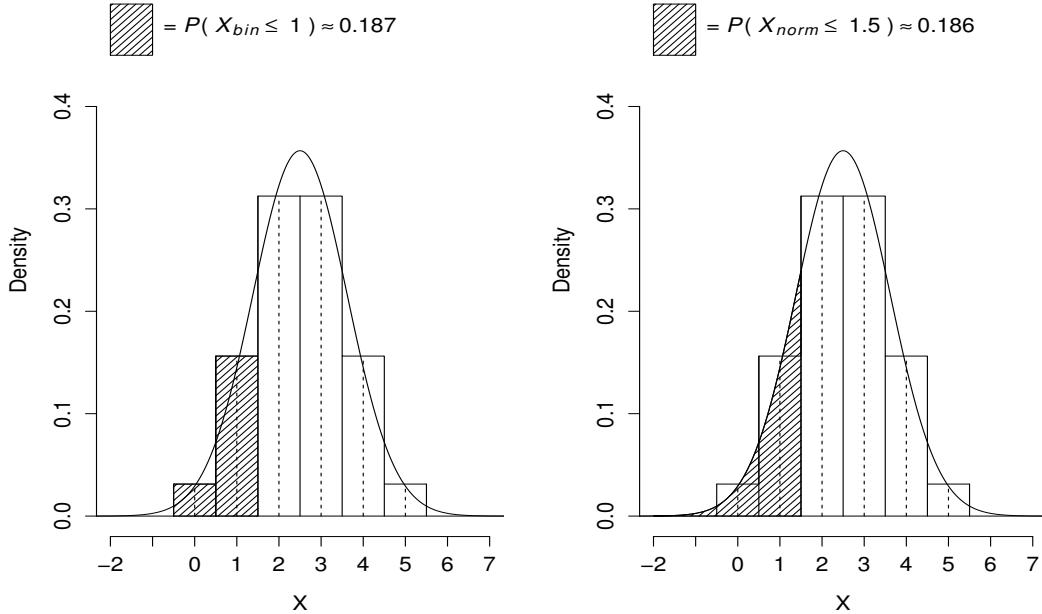


Figure 4.10: Histogram representation of the distribution of  $X_{bin} \sim bin(5, 1/2)$ , with the normal density for  $X_{norm} \sim N(5(1/2), 5(1/2)^2)$  superimposed. The left plot illustrates the calculation of  $P(X_{bin} \leq 1)$ , while the right plot illustrates the calculation of  $P(X_{norm} \leq 1.5)$ .

## 4.15 The $\chi^2$ , $t$ - and $F$ - distributions

There are a number of distributions which are essential to statistical inference, playing a role in many procedures. They are based on the normal distribution.

### 4.15.1 The $\chi^2$ Distribution

Suppose  $X_1, \dots, X_n$  is an *iid* sample from standard normal distribution  $N(0, 1)$ . If we let

$$W = \sum_{i=1}^n X_i^2$$

then we say  $W$  possesses a  $\chi_n^2$  distribution with  $n$  *degrees of freedom*. Figure 4.11 gives examples of the  $\chi^2$  density for 5, 10 and 20 degrees of freedom.

As for the normal distribution, probability values are tabulated. However, because there is a separate density for each degree of freedom, the available probabilities are much more limited. Table A.4 give critical values

$x_\alpha$  for commonly used values of  $\alpha$  (Appendix A). The use of these tables will be discussed in more detail in Chapter 18.

#### 4.15.2 The $F$ -distribution

Suppose  $W_1 \sim \chi^2_{\nu_1}$  and  $W_2 \sim \chi^2_{\nu_2}$ , and that  $W_1$  and  $W_2$  are independent. If we set

$$F = \frac{W_1/\nu_1}{W_2/\nu_2},$$

then we say  $F$  possesses an  $F_{\nu_1, \nu_2}$  distribution with  $\nu_1$  numerator degrees of freedom and  $\nu_2$  denominator degrees of freedom. Figure 4.11 gives examples of the  $F$  density for  $(5, 5)$ ,  $(10, 20)$  and  $(2, 50)$  numerator and denominator degrees of freedom.

Tables A.5-A.8 give critical values  $x_\alpha$  for commonly used values of  $\alpha$  (Appendix A). The use of these tables will be discussed in more detail in Chapter 18.

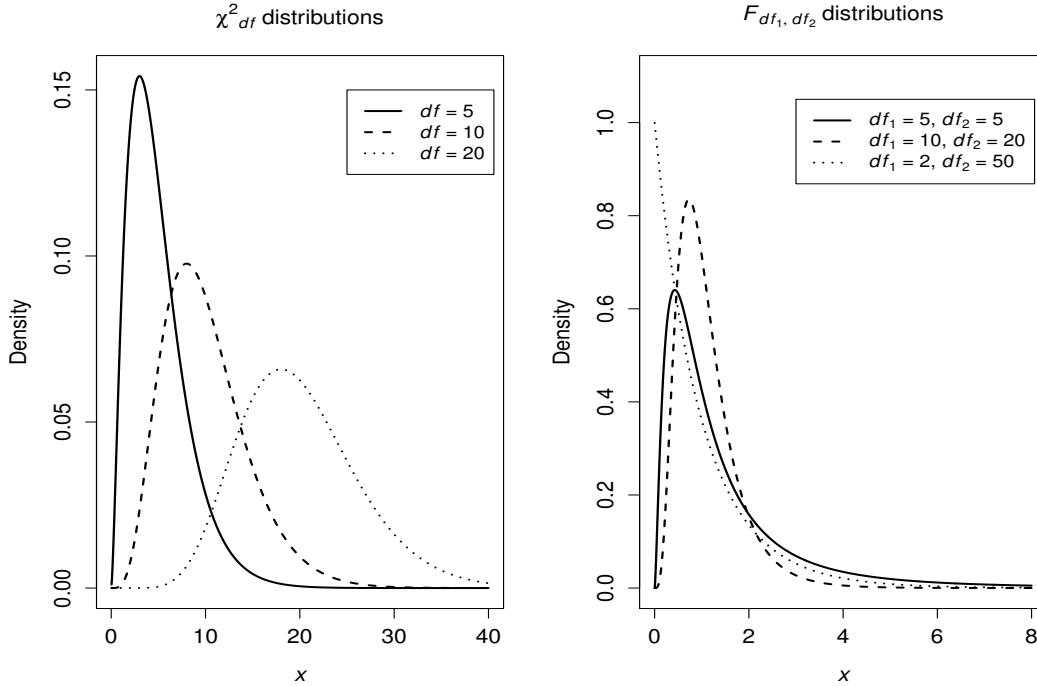


Figure 4.11: Examples of  $\chi^2_{df}$  densities and  $F_{df_1, df_2}$  densities.

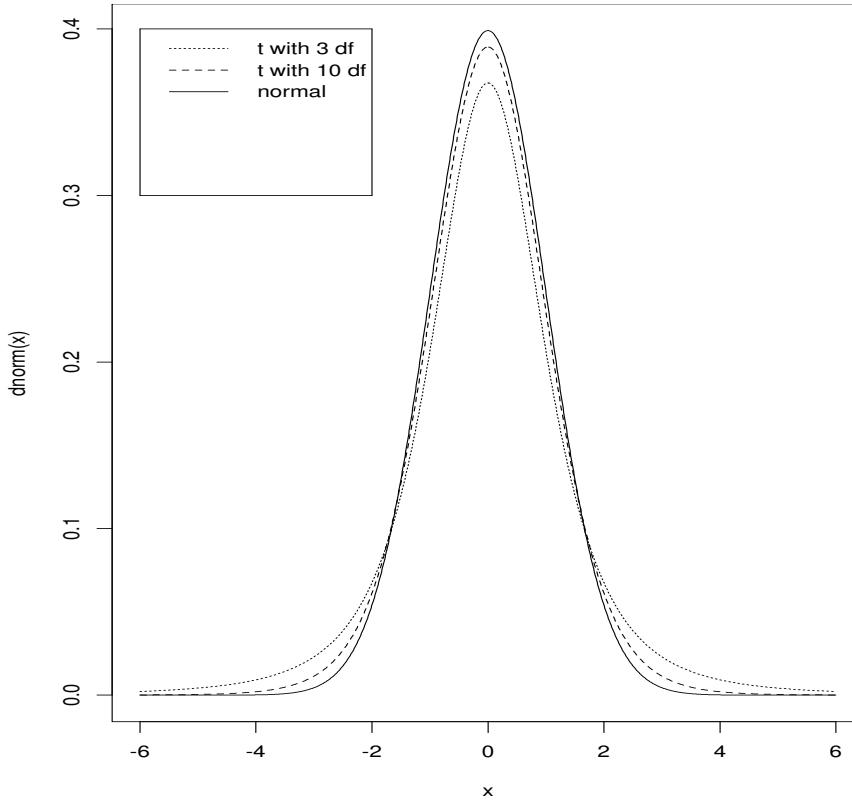


Figure 4.12: Several  $t$ -distribution density functions

#### 4.15.3 The $t$ -distribution

Formally, the  $t$ -distribution with  $\nu$  degrees of freedom is equivalent to

$$T = \frac{Z}{\sqrt{W/\nu}} \quad (4.35)$$

where  $Z \sim N(0, 1)$ ,  $W \sim \chi_{\nu}^2$  and  $Z \perp W$ .

As we might expect, the  $t$ -distribution resembles a normal distribution, except that there is somewhat more variability. As the degrees of freedom increase, the resemblance to the normal becomes greater. Note that the last line of the table is labeled as  $df = \infty$ . This corresponds to the standard normal distribution. Figure 4.12 plots several densities of the  $t$ -distribution.

Table A.3 gives critical values  $x_{\alpha}$  for commonly used values of  $\alpha$  (Appendix A). The use of these tables will be discussed in more detail in Section 12.3.

#### 4.16 Survival Times

A *survival time*  $T$  is a nonnegative random variable variously interpreted as a *survival time*, *lifetime*, *waiting time*, *time to event*, and so on. The

geometric and exponential RVs (Sections 4.10.4 and 4.11.2) are two examples of survival times, so a survival time may be discrete (for example, number of integer days until event) or continuous. It might seem counterintuitive that a survival time may be memoryless, precisely the property which characterizes the geometric and exponential distributions. We usually don't expect that the average waiting time remaining, after having already waited a time  $t$ , should be the same as the original average waiting time.

The key to understanding this idea is to consider a *failure rate*. Suppose a survival time  $T$  is discrete, representing the number of days until failure of a component (a light bulb, for example). Let  $q_i$  be the failure rate on day  $i$ . This means that if the component has survived until day  $i$  ( $T \geq i$ ) we toss a coin (independently of all previous coin tosses). If we get a ‘head’ (for our particular coins, this has probability  $q_i$ ) the component ‘fails’ and the survival time is  $T = i$ , terminating the process. First, note that the failure rates  $q_i$  are not a PMF for  $T$ . They are actually the conditional probabilities

$$q_i = P(T = i \mid T \geq i), \quad i = 1, 2, \dots$$

Second, these rates may increase or decrease in time, and what defines a memoryless distribution is precisely the assumption that the failure rates remain constant. This defines the geometric distribution.

We do not, on the other hand, expect the lifetime of, say, a car to be memoryless. We expect that the probability that a 10-year old car survives one more year is smaller than that for a 5-year old car, and smaller still than that for a new car. In other words, the failure rates  $q_i$  increase in  $i$ . Such a survival time is called *new better than used (NBU)*.

A survival time may also be *new worse than used (NWU)*, in which case the failure rates are decreasing. The survival time for young members of a species in an environment with high infant mortality will typically be NWU. This is because the period immediately after birth is very high in mortality risk, meaning that the failure rate is correspondingly high. However, if the infant survives this high risk period, the failure rate will decrease, resulting in a NWU survival time. Of course, if the infant survives into adulthood, the failure rate will begin to increase. A natural source of NWU survival times would be survival in competitive environments.

## 4.17 Power Law Distributions

A *power law distribution* is a discrete distribution on sample space  $S = \{1, 2, \dots, M\}$  for which the PMF is proportional to

$$p_X(k) \propto \frac{1}{k^\alpha}, \quad k = 1, \dots, M, \tag{4.36}$$

for some positive constant  $\alpha > 0$ . The PMF can be normalized in a straightforward manner,

$$p_X(k) = \frac{c}{k^\alpha} = \frac{k^{-\alpha}}{\sum_{i=1}^M i^{-\alpha}}, \quad k = 1, \dots, M, \quad (4.37)$$

where

$$c^{-1} = \sum_{i=1}^M i^{-\alpha},$$

forcing the PMF to sum to 1. Note that when using this convention, we have  $p_X(1) = c$ .

Note that the probabilities  $p_X(k)$  are decreasing in  $k$ . Sometimes, a power law is created from another discrete distribution by sorting the sample space in decreasing order of frequency, so that  $k = 1$  is the most frequent outcome,  $k = 2$  is the second most frequent outcome, and so on. We can assume in this section that this has been done if needed.

*Zipf's law* is a power law distribution in which  $\alpha = 1$  in (4.36)-(4.37). It is named after the linguist George Zipf (1902-1950), who reported that the frequency of words in a natural language conformed to this distribution, in particular, the most common word tends to appear twice as often as the second most common, three times as often as the third most common word, and so on. The power law distribution has assumed increasing prominence with the greater availability of data sets of a quite rich variety, and appears to occur quite naturally in a large variety of processes. Empirical comparisons similar to those earlier computed for word frequencies are frequently reported.

The comparison is easily made using a *log-log* plot. To see this, take a *log* transform of each side of (4.37), which gives

$$\begin{aligned} \log(p_X(k)) &= \log(c) - \alpha \log(k) \\ &= \log(p_X(k)) \\ &= \log(p_X(1)) - \alpha \log(k), \quad k = 1, \dots, M. \end{aligned} \quad (4.38)$$

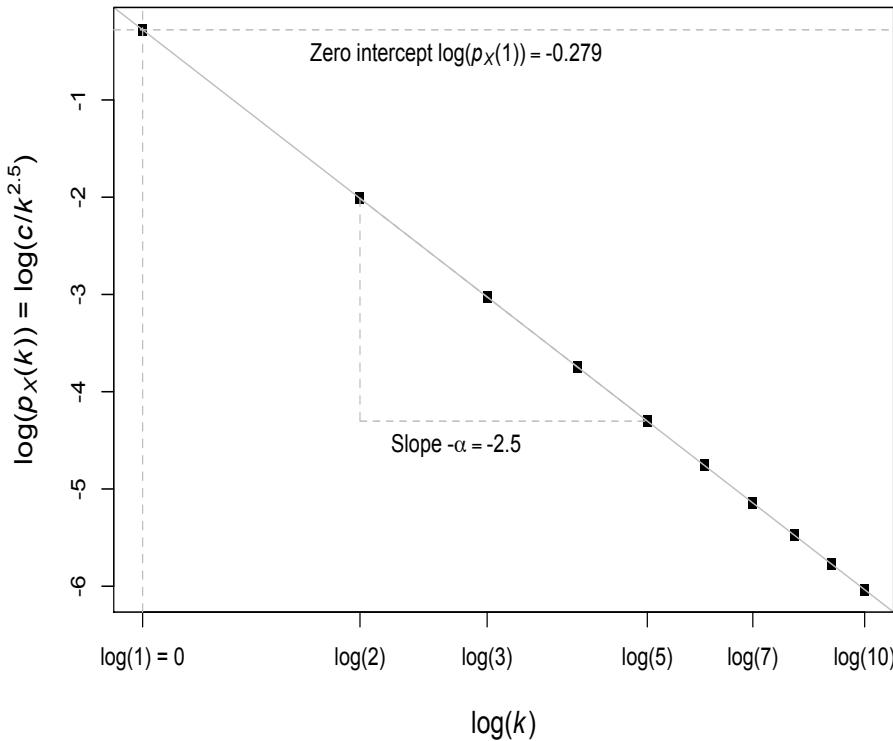
Then, plot the points  $(\log(k), \log(p_X(k)))$  on a graph. For an empirical comparison  $p_X(k)$  represents the *observed* or *empirical frequencies*. If the frequencies conform to the power law, the plot should be approximately a straight line, with slope  $-\alpha$ , and zero-intercept  $\log(p_X(1))$  (that is, the value at the vertical axis when the line crosses 0 at the horizontal axis).

**Example 4.28** Consider a power law distribution (4.37) with  $M = 10$  and  $\alpha = 2.5$ . This gives, from (4.37),

$$p_X(k) = \frac{c}{k^{2.5}}, \quad k = 1, \dots, 10$$

where

$$c^{-1} = \sum_{i=1}^{10} i^{-2.5} = \frac{1}{1} + \frac{1}{2^{2.5}} + \dots + \frac{1}{10^{2.5}} \approx 1.322 \approx 0.756^{-1}.$$

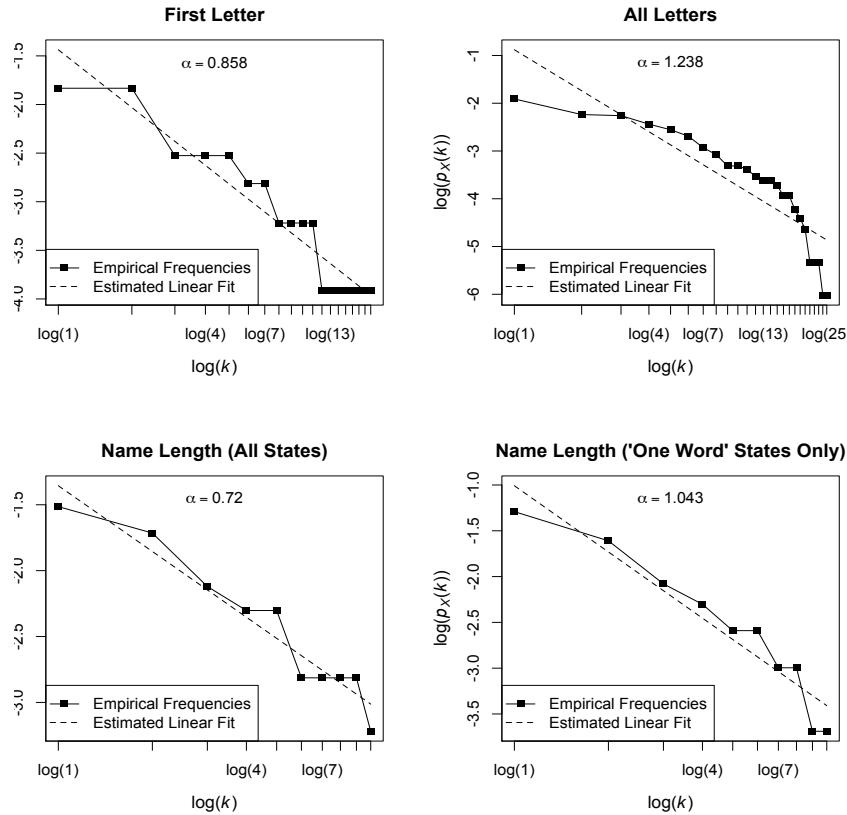
*log-log plot of power law distribution with  $M = 10$  and  $\alpha = 2.5$* Figure 4.13: *Log-log plot of power law distribution with  $M = 10$ ,  $\alpha = 2.5$ , for Example 4.28.*

This gives

$$p_X(k) \approx \frac{0.756}{k^{2.5}}, \quad k = 1, \dots, 10,$$

so that, for example,  $p_X(1) \approx 0.756$ ,  $p_X(2) \approx 0.756/2^{2.5} = 0.134$ , and so on. Figure 4.13 shows a *log-log* plot of the PMF. Because the distribution is given exactly, we expect the linear relation (4.38) to hold exactly. As expected, the slope is  $-\alpha = -2.5$  and the zero-intercept is  $\log(p_X(1)) = \log(0.756) = -0.279$ . ■

We next consider an empirical comparison.

Figure 4.14: *Log-log* plots for Example 4.29.

**Example 4.29** Various empirical distributions can be formed from the names of the 50 states. We consider the following:

- A The first letters of the names.
- B All letters from the names.
- C The length of the names.
- D The length of ‘one word’ names.

In all cases blanks are excluded. In case D ‘two word’ states such as ‘New York’ are excluded. For example, the frequencies for case A is given below:

N	M	W	I	A	O	C	V	T	S	K	U	R	P	L	H	G	F	D
8	8	4	4	4	3	3	2	2	2	2	1	1	1	1	1	1	1	

so that 8 state names begin with letter ‘N’ and one state name begins with letter ‘D’. It is difficult to determine by examining these frequencies in this form whether or not they conform to the Zipf law. In Figure 4.14 *log-log* plots are drawn for each case. A linear fit is superimposed (this will be

discussed in Chapter 25), resulting in an estimate of the power  $\alpha$  in (4.37). We find that the Zipf law is at least plausible for cases  $A$ ,  $C$  and  $D$  but not  $B$ . Interestingly, the Zipf law appears more plausible for case  $D$  than for case  $C$ , that is, after the ‘two word’ states have been removed. Possibly, allowing ‘two word’ names contaminates the distribution, in the sense that any distributional process would depend on the number of words permitted. This would result in the pooling of two distinct distributions, which tends to complicate this type of analysis. ■

An excellent survey of this topic can be found in Clauset *et al.* (2009).

## 4.18 Random Variables in R

R contains functions associated with a wide class of random variables. There exists a standard naming convention. For example, the normal distribution has R name `norm`, and the binomial distribution has R name `binom`. For each such name there are four functions with prefix ‘d’ for density, ‘p’ for CDF, ‘q’ for quantile and ‘r’ for simulated random variate. As needed, each R name has additional arguments needed to completely specify the distribution. For the binomial those are `size` and `prob` corresponding to `n,p` in the convention  $X \sim bin(n, p)$ . For example, for  $X \sim bin(10, 0.25)$  the commands

```
> pbinom(3, size=10, prob=0.25)
[1] 0.7758751
> dbinom(3, size=10, prob=0.25)
[1] 0.2502823
> qbinom(0.5, size=10, prob=0.25)
[1] 2
>
```

tell us that  $P(X \leq 3) = 0.7758751$ ,  $P(X = 3) = 0.2502823$  and that 2 is (approximately) the median. We can simulate a random sample of size `n` in the following way:

```
> n = 5
> x = rbinom(n, size=10, prob=0.25)
> x
[1] 2 4 0 2 5
```

The following list is part of the R documentation. See also Chapter 8 of *An Introduction to R* for a complete list of name conventions.

For the beta distribution see `dbeta`.

For the binomial (including Bernoulli) distribution see `dbinom`.

For the Cauchy distribution see `dcauchy`.

For the chi-squared distribution see `dchisq`.

For the exponential distribution see `dexp`.  
For the F distribution see `df`.  
For the gamma distribution see `dgamma`.  
For the geometric distribution see `dgeom`. (This is also ...)  
For the hypergeometric distribution see `dhyper`.  
For the log-normal distribution see `dlnorm`.  
For the multinomial distribution see `dmultinom`.  
For the negative binomial distribution see `dnbinom`.  
For the normal distribution see `dnorm`.  
For the Poisson distribution see `dpois`.  
For the Student's t distribution see `dt`.  
For the uniform distribution see `dunif`.  
For the Weibull distribution see `dweibull`.

For less common distributions of test statistics see ...

#### See Also

RNG about random number generation in R.

## Chapter 5

# Bayes' Theorem and Classification

If we are given a conditional probability  $P(E | A)$  we often would like to “reverse the order” of the events to obtain  $P(A | E)$ . To do this we use *Bayes’ theorem*

**Theorem 5.1** For two events  $A$  and  $E$ , with  $P(E) > 0$ , we have

$$P(A | E) = P(E | A) \frac{P(A)}{P(E)}. \quad (5.1)$$

This is referred to as Bayes’ theorem. ■

*Proof.* The Equation (5.1) is proven with the following argument:

$$\begin{aligned} P(A | E) &= \frac{P(AE)}{P(E)} \\ &= \frac{P(E | A)P(A)}{P(E)}. \end{aligned} \quad \square$$

The following definition, though straightforward, is quite important to understanding the current chapter.

**Definition 5.1** In the context of Baye’s theorem  $P(A)$  is the *prior probability* of  $A$ , and  $P(A | E)$  is the *posterior probability* of  $A$  given information  $E$ . ■

The following is a useful variation of Baye’s theorem.

**Theorem 5.2** Suppose events  $A_1, \dots, A_n$  is a partition of sample space  $S$ , that is, the events are mutually exclusive with

$$S = \bigcup_{i=1}^n A_i.$$

For any  $1 \leq i \leq n$

$$\begin{aligned} P(A_i | E) &= \frac{P(E | A_i)P(A_i)}{P(E)} \\ &= \frac{P(E | A_i)P(A_i)}{P(E | A_1)P(A_1) + \dots + P(E | A_n)P(A_n)}. \end{aligned}$$

*Proof.* The theorem follows from the *law of total probability* given in Definition 2.10.  $\square$

**Example 5.1** Suppose a test for a certain infection is evaluated by administering the test to 50 patients with the infection, and 100 patients known to be without the infection (control patients). The test was positive for 49 of the 50 infected patients and positive for 4 of the 100 control patients. Let

$$\begin{aligned} T &= \{ \text{Patient tests positive} \} \\ D &= \{ \text{Patient has infection} \} \end{aligned}$$

From the above data we can estimate directly

$$\begin{aligned} P(T | D) &= 49/50 \\ P(T^c | D) &= 1/50 \\ P(T | D^c) &= 4/100 \\ P(T^c | D^c) &= 96/100. \end{aligned}$$

Thus, from the data we get directly

$$P(T | D) = \text{Probability of testing positive when infected}$$

and

$$P(T | D^c) = \text{Probability of testing positive when not infected}$$

but what is of ultimate interest are the probabilities

$$P(D | T) = \text{Probability of being infected when testing positive}$$

and

$$P(D | T^c) = \text{Probability of being infected when not testing positive}$$

since this is the quantity which is clinically relevant. We use Baye's Theorem to calculate these probabilities, setting

$$\begin{aligned} A_1 &= D \\ A_2 &= D^c. \end{aligned}$$

Then

$$\begin{aligned} P(D | T) &= \frac{P(T | D)P(D)}{P(T | D)P(D) + P(T | D^c)P(D^c)} \\ &= \frac{(49/50)P(D)}{(49/50)P(D) + (4/100)P(D^c)} \end{aligned}$$

and

$$\begin{aligned} P(D | T^c) &= \frac{P(T^c | D)P(D)}{P(T^c | D)P(D) + P(T^c | D^c)P(D^c)} \\ &= \frac{(1/50)P(D)}{(1/50)P(D) + (96/100)P(D^c)}. \end{aligned}$$

Note that in order to evaluate these probabilities we need to know  $P(D)$  which was not obtained from the original experiment. This should be the case, since if  $P(D) = 0$  (*i.e.*, the infection is nonexistent) we would also expect both  $P(D | T) = P(D | T^c) = 0$ . ■

## 5.1 Odds

The term *odds* is synonymous with probability, and is formally defined as follows:

**Definition 5.2** For a given event  $A$  we define the *odds* to be

$$Odds(A) = \frac{P(A)}{P(A^c)} = \frac{P(A)}{1 - P(A)}. \quad \blacksquare$$

If we roll a die, the probability of getting a six is  $1/6$ , but the odds are  $1/5$ . Mathematically, the odds and the probability of an event  $A$  are equivalent, since we can calculate the odds from the probability, as well as the probability from the odds:

$$P(A) = \frac{Odds(A)}{1 + Odds(A)}.$$

In particular, if  $A$  is certain to occur then

$$\begin{aligned} P(A) &= 1 \\ Odds(A) &= \infty \end{aligned}$$

and if  $A$  is certain to not occur then

$$\begin{aligned} P(A) &= 0 \\ Odds(A) &= 0. \end{aligned}$$

We can also define the *conditional odds* of  $A$  given  $E$ .

**Definition 5.3** The *conditional odds* of  $A$  given  $E$  is defined as

$$Odds(A | E) = \frac{P(A | E)}{P(A^c | E)} = \frac{P(A | E)}{1 - P(A | E)}. \quad \blacksquare$$

The conditional odds leads to a particularly intuitive form of Baye's theorem.

**Theorem 5.3** The conditional odds of  $A$  given  $E$  may be expressed

$$Odds(A | E) = \frac{P(E | A)}{P(E | A^c)} \times Odds(A). \quad (5.2)$$

assuming  $P(E) > 0$ . ■

*Proof.* Equation (5.2) is proven with the following argument:

$$\begin{aligned} Odds(A | E) &= \frac{P(A | E)}{P(A^c | E)} \\ &= \frac{P(E | A)P(A)}{P(E)} \times \frac{P(E)}{P(E | A^c)P(A^c)} \\ &= \frac{P(E | A)}{P(E | A^c)} \times \frac{P(A)}{P(A^c)} \\ &= \frac{P(E | A)}{P(E | A^c)} \times Odds(A). \end{aligned} \quad \square$$

## 5.2 The Bayesian Model

Under the **Bayesian model** we are interested in the probability of a hypothesis  $A$ , or more specifically, the effect on this probability of the introduction of information or evidence  $E$ . There may be a well known prevalence of a certain condition (hypothesis  $A$ ) among a population. For any given patient entering a clinic, this prevalence may be  $P(A)$ . A diagnostic test is then done. Let  $E$  be the event that this test is positive. We are now no longer interested in  $P(A)$ , but in  $P(A | E)$  or  $P(A | E^c)$ , depending on the outcome of the test.

Based on an evaluation of the accuracy of the test, we may know  $P(E | A)$  and  $P(E | A^c)$ . Examining Equation (5.2), we define the *likelihood ratio* as follows:

**Definition 5.4** When considering the odds of an event  $A$  given evidence  $E$ , the *likelihood ratio* is given by

$$LR = \frac{P(E | A)}{P(E | A^c)},$$

from which we get, as a reexpression of Theorem 5.3,

$$Odds(A | E) = LR \times Odds(A). \quad (5.3)$$

We refer to  $Odds(A)$  as the *prior odds* and to  $Odds(A | E)$  as the *posterior odds*. ■

The relationship between the prior and posterior odds is the same as that between the prior and posterior probability. However, Equation (5.3) very neatly captures the ability of the evidence to alter our assessment of the probability of a hypothesis in a way which does not depend on the prior probability.

**Example 5.2** We will express the previous Example 5.1 in terms of odds of having the infection. If a patient tests positive, the odds are adjusted by the formula

$$\begin{aligned} Odds(D | T) &= \frac{P(T | D)}{P(T | D^c)} \times Odds(D) \\ &= \frac{49/50}{4/100} \times Odds(D) \\ &= 24.5 \times Odds(D) \end{aligned}$$

so that testing positive *increases* the odds of having the infection by a factor of 24.5.

If the patient tests negative, the odds are adjusted by the formula

$$\begin{aligned} Odds(D | T^c) &= \frac{P(T^c | D)}{P(T^c | D^c)} \times Odds(D) \\ &= \frac{1/50}{96/100} \times Odds(D) \\ &= \frac{1}{48} \times Odds(D) \end{aligned}$$

so that testing negative *decreases* the odds of having the infection by a factor of 48.

We are therefore in a better position to evaluate the accuracy of the test when the problem is expressed in terms of odds. ■

**Example 5.3** Suppose blood collected at a crime scene is typed for DNA. A genotype if found which is estimated to occur in the population with a frequency of  $p$ . A suspect is similarly typed and found to have the same genotype. Suppose

$$\begin{aligned} A &= \{ \text{Suspect's blood is that found at the crime scene} \} \\ E &= \{ \text{Suspect has the same genotype as blood found at crime scene} \} \end{aligned}$$

Then the likelihood ratio is constructed by noting that

$$\begin{aligned} P(E | A) &= 1 \\ P(E | A^c) &= p \end{aligned}$$

giving

$$\begin{aligned} LR &= \frac{P(E | A)}{P(E | A^c)} \\ &= \frac{1}{p} \end{aligned}$$

so that the odds that the blood is the same is adjusted by

$$\begin{aligned} Odds(A | E) &= LR \times Odds(A) \\ &= \frac{1}{p} \times Odds(A) \end{aligned}$$

We usually have no way of directly evaluating  $Odds(A)$ . We can only describe how the evidence changes the odds. If it were established without doubt that the suspect was not at the crime scene by other evidence then we would have

$$Odds(A) = 0$$

and

$$Odds(A | E) = 0$$

for any value of  $LR$ . If guilt were established with absolute certainty then

$$Odds(A) = \infty$$

and

$$Odds(A | E) = \infty.$$

for any value of  $LR$ .

Now, suppose the genotype does not match. (That is,  $E^c$  occurs). The likelihood ratio is now calculated from

$$\begin{aligned} P(E^c | A) &= 0 \\ P(E^c | A^c) &= 1 - p \end{aligned}$$

giving  $LR = 0$  so that

$$\begin{aligned} Odds(A | E) &= LR \times Odds(A) \\ &= 0 \end{aligned}$$

for any  $Odds(A)$ . ■

### 5.3 The Fallacy of the Transposed Conditional

In the previous example suppose we set  $p = 1/100$ . We could then say

$$P(E | A^c) = 1/100$$

which is the probability of a genotype match if the suspect is not guilty. A common error is to transpose the conditional which yields (incorrectly)

$$P(A^c | E) = 1/100.$$

This statement says that the probability that the suspect is not guilty is 1/100 if a match occurs. After some algebra we then get

$$\begin{aligned} P(A | E) &= 1 - P(A^c | E) \\ &= 99/100 \end{aligned}$$

which says that, given a match, the probability that the suspect is guilty is 99/100, which cannot be concluded from the evidence. The odds of guilt given the evidence depends on the prior odds. This is often referred to as the *prosecutor's fallacy*.

## 5.4 Diagnostic Testing - Basic Definitions

A common problem in medical research is the evaluation of the accuracy of diagnostic tests. This can be framed in the context of probability theory. In the simplest case, a diagnostic test is either positive (in which case the patient is predicted to have the condition being tested) or negative (in which case the patient is predicted to not have the condition being tested). There are 4 events of interest:

$$\begin{aligned} O_- &= \{ \text{the patient does not have the condition} \} \\ O_+ &= \{ \text{the patient has the condition} \} \\ T_- &= \{ \text{the patient tests negative} \} \\ T_+ &= \{ \text{the patient tests positive} \} \end{aligned}$$

Clearly,  $O_-^c = O_+$  and  $T_-^c = T_+$ , so that  $P(O_-) + P(O_+) = 1$  and  $P(T_-) + P(T_+) = 1$ .

Conditional probabilities and Bayes theorem can be very useful in developing a probabilistic model for these outcomes, and a widely used terminology has been developed:

$$\begin{aligned} \text{sensitivity (sens)} &= P(T_+ | O_+) \\ \text{specificity (spec)} &= P(T_- | O_-) \\ \text{positive predictive value (PPV)} &= P(O_+ | T_+) \\ \text{negative predictive value (NPV)} &= P(O_- | T_-) \\ \text{prevalance (prev)} &= P(O_+). \end{aligned} \tag{5.4}$$

Two more related definitions are sometimes used:

$$\begin{aligned} \text{true discovery rate (TDR)} &= \text{sens} \\ \text{false discovery rate (FDR)} &= P(T_+ | O_-) = 1 - \text{spec}. \end{aligned}$$

In an evaluation study, a diagnostic test will typically be administered to subjects with known outcomes, which will allow sensitivity and specificity to be estimated. However, when used in a clinical setting, the outcomes

will not be known. These are to be predicted based on the test result, so it will be  $PPV$  and  $NPV$  which are more relevant. These quantities can be related to sensitivity, specificity and prevalence using Baye's Theorem:

$$\begin{aligned} PPV &= P(O_+ | T_+) \\ &= \frac{P(T_+ | O_+)P(O_+)}{P(T_+ | O_+)P(O_+) + P(T_+ | O_-)P(O_-)} \\ &= \frac{sens \times prev}{sens \times prev + (1 - spec) \times (1 - prev)} \end{aligned} \quad (5.5)$$

and

$$\begin{aligned} NPV &= P(O_- | T_-) \\ &= \frac{P(T_- | O_-)P(O_-)}{P(T_- | O_-)P(O_-) + P(T_- | O_+)P(O_+)} \\ &= \frac{spec \times (1 - prev)}{spec \times (1 - prev) + (1 - sens) \times prev}. \end{aligned} \quad (5.6)$$

It is important to understand the degree to which  $PPV$  and  $NPV$  depend on prevalence. We have already seen in Example 5.1 that if, for example,  $prev = 0$  we would necessarily have  $PPV = 0$ , no matter what sensitivity and specificity are. On the other hand, sensitivity and specificity do not depend on prevalence, and this distinction is an important one.

#### 5.4.1 Diagnostic Tests and Contingency Tables

The outcomes of a study used to evaluate a diagnostic test can be summarized in Table 5.1 below,

Table 5.1: Outcomes of diagnostic testing

		Condition	
		Positive	Negative
Test	Positive	TP	FP
	Negative	FN	TN

where

$$\begin{aligned} TP &= T_+ \cap O_+ = \text{True Positive} \\ FP &= T_+ \cap O_- = \text{False Positive} \\ TN &= T_- \cap O_- = \text{True Negative} \\ FN &= T_- \cap O_+ = \text{False Negative.} \end{aligned} \quad (5.7)$$

Table 5.1 can be interpreted as a contingency table, with numerical entries TP, FP, TN, FN giving the counts of subjects in each category. These can be

used to estimate all important quantities. If we let  $N = TP + FP + TN + FN$  (the total number of entries in Table 5.1) we can calculate the *marginal probabilities*:

$$\begin{aligned} P(O_-) &= \frac{FP + TN}{N} \\ P(O_+) &= \frac{TP + FN}{N} \\ P(T_-) &= \frac{FN + TN}{N} \\ P(T_+) &= \frac{TP + FP}{N}, \end{aligned} \tag{5.8}$$

and the important diagnostic quantities

$$\begin{aligned} prev &= \frac{TP + FN}{N} = P(O_+) \\ sens &= \frac{TP}{TP + FN} = P(T_+ | O_+) \\ spec &= \frac{TN}{TN + FP} = P(T_- | O_-) \\ PPV &= \frac{TP}{TP + FP} = P(O_+ | T_+) \\ NPV &= \frac{TN}{TN + FN} = P(O_- | T_-). \end{aligned} \tag{5.9}$$

However, the prevalence must be very carefully interpreted. If we calculate *prev* directly from Table 5.1, we obtain the prevalence of an outcome within the study population, which may have no relationship to the prevalence in any given clinical population. This would be especially true if the study was designed to ensure a large enough sample of disease positive subjects to accurately evaluate the test. In such cases, we would expect *prev* to be much higher than it would be in a clinical population.

Therefore, it is important to understand that it is always possible, and usually preferable, to calculate prevalence independently of sensitivity and specificity. In particular, if we are using a study such as that represented by Table 5.1 we would use equations (5.9) to estimate *sens* and *spec* but not *prev*, *PPV* or *NPV*. Instead, we would use an independent estimate of *prev* which more accurately estimates the prevalence within the clinical population of interest, and then use (5.5)-(5.6) with that value of *prev*.

In summary, the important question is whether or not the subjects used in Table 5.1 are representative of the population in which the test is to be applied, in terms of the relative frequencies of outcomes  $O_+$  and  $O_-$ . The values of *prev*, *PPV* and *NPV* calculated by equations (5.9) would be interpretable only if this is the case.

### 5.4.2 The Use of Odds in the Evaluation of Diagnostic Tests

In the absence of a reliable estimate of prevalence, the accuracy of a diagnostic test can be expressed using odds, as shown above. Using the previous terminology we have

$$LR = \frac{P(T_+ | O_+)}{P(T_+ | O_-)} = \frac{\text{sensitivity}}{1 - \text{specificity}}$$

so that

$$Odds(O_+ | T_+) = LR \times Odds(O_+).$$

Then  $Odds(O_+)$  is the prevalence expressed as odds, and the predictive ability of the test can be expressed using only the sensitivity and specificity.

Note that we can also assess the accuracy of a negative test outcome. In this case we can distinguish between the  $LR$  for a positive test outcome  $LR_+$  and the  $LR$  for a negative test outcome  $LR_-$ :

$$LR_+ = LR \text{ as defined above } LR_- = \frac{P(T_- | O_+)}{P(T_- | O_-)} = \frac{1 - \text{sensitivity}}{\text{specificity}}$$

so that

$$\begin{aligned} Odds(O_+ | T_+) &= LR_+ \times Odds(O_+) \\ Odds(O_+ | T_-) &= LR_- \times Odds(O_+). \end{aligned}$$

**Example 5.4** Studies into the accuracy of a diagnostic test often proceed by pairing the test with a *gold standard* in a study group of size  $N$ , the latter assumed to be perfectly accurate. In this case, we can estimate sensitivity and specificity. After the study we would construct a contingency table like the following ( $N = 1000$ ):

Table 5.2: Outcomes of diagnostic testing for Example 5.4

	Condition		
	Positive	Negative	
Test	Positive	30	110
	Negative	10	850

We have, using equations (5.9), ( $TP = 30$ , for example):

$$\begin{aligned} sens &= 30/40 = 0.75 \\ spec &= 850/960 \approx 0.885 \\ LR_+ &= (30/40)/(1 - 850/960) \approx 6.545 \\ LR_- &= (1 - 30/40)/(850/960) \approx 0.282 \end{aligned}$$

and the Bayes model gives

$$\begin{aligned} Odds(O_+ | T_+) &\approx 6.545 \times Odds(O_+) \\ Odds(O_+ | T_-) &\approx 0.282 \times Odds(O_+). \end{aligned}$$

A positive test result increases the odds of a positive outcome by a factor of 6.545, while a negative test result decreases the odds of a positive outcome by a factor of 0.282.

Next, if we calculate  $prev$ ,  $PPV$  and  $NPV$  directly from the contingency table, using equations (5.9). we would have

$$\begin{aligned} prev &= (10 + 30)/1000 = 0.04 \\ PPV &= 30/140 \approx 0.214 \\ NPV &= 850/860 \approx 0.988. \end{aligned}$$

The values of  $PPV$  and  $NPV$  assume a prevalence of 4%, estimated directly from the data. Suppose the true prevalence was 2%. We would then use (5.5)-(5.6) with  $prev = 0.02$  and the estimates of  $sens$  and  $spec$  obtained from the data (remember that these quantities do not depend on the prevalence). This gives

$$\begin{aligned} PPV &= \frac{sens \times prev}{sens \times prev + (1 - spec) \times (1 - prev)} \\ &= \frac{0.75 \times 0.02}{0.75 \times 0.02 + (1 - 0.885) \times (1 - 0.02)} \\ &\approx 0.117 \end{aligned}$$

and

$$\begin{aligned} NPV &= \frac{spec \times (1 - prev)}{spec \times (1 - prev) + (1 - sens) \times prev} \\ &= \frac{0.885 \times (1 - 0.02)}{0.885 \times (1 - 0.02) + (1 - 0.75) \times 0.02} \\ &\approx 0.994. \end{aligned}$$

Reducing the prevalence by 1/2 results in a reduction in  $PPV$  of almost the same magnitude (verify that if we use  $prev = 0.04$  in equations (5.5)-(5.6) we reproduce the values of  $PPV$  and  $NPV$  obtained using equations (5.9)).

Using either method, that  $PPV$  is much smaller than sensitivity is typical, and is due to the fact that  $PPV$  depends on the prevalence. Expecting the two to be equal is an example of the ‘prosecutor’s fallacy’, since one is obtained from the other by transposing the conditional.

Note also that  $NPV$  is quite large. This is a function both of the ability of the test to rule out a positive outcome (measured by specificity) and of the relatively small prevalence. This means  $NPV$  would be smaller when the test is confined to a higher risk population. ■

## 5.5 The Odds Ratio

Consider the following events

$$\begin{aligned} O_- &= \{ \text{the patient does not have the condition} \} \\ O_+ &= \{ \text{the patient has the condition} \} \\ G_1 &= \{ \text{the patient is in Group 1} \} \\ G_2 &= \{ \text{the patient is in Group 2} \}. \end{aligned}$$

Typically, we are interested in comparing

$$P(O_+ | G_1) \text{ and } P(O_+ | G_2).$$

We will discuss this topic further in Chapter 15, but it is worth considering the problem here briefly, in the context of this chapter.

Perhaps the obvious comparison method is to examine the difference:

$$\Delta = P(O_+ | G_1) - P(O_+ | G_2).$$

This will be, sometimes, a reasonable approach, but will not work well when the probabilities are small. Alternatively, we have the *relative risk*

$$RR = \frac{P(O_+ | G_1)}{P(O_+ | G_2)}$$

and the *odds ratio* (OR)

$$OR = \frac{Odds(O_+ | G_1)}{Odds(O_+ | G_2)} = \frac{P(O_+ | G_1)/(1 - P(O_+ | G_1))}{P(O_+ | G_2)/(1 - P(O_+ | G_2))}.$$

The OR has an interesting property, in that events defining it may be transposed, that is

$$OR = \frac{Odds(G_1 | O_+)}{Odds(G_1 | O_-)},$$

so that the OR does not depend on the marginal probabilities (that is, the prevalences). For some applications, this is a considerable advantage, for the reasons discussed earlier in this chapter.

## Chapter 6

# Simulation of Random Variables

The simulation of random variables is an indispensable tool for statistics, and many other branches of applied mathematics. Many computing platforms include random variable simulation functions, including R (see Section 4.18), and in this chapter we review some of the relevant simulation methods. Knowledge of these methods is important for a number of practical reasons. First, this is needed to understand the issue of *reproducibility*, that is, the ability of a second analyst to reproduce independently the results of a first analyst. The use of simulated random variables in an analysis is especially relevant to this issue. Second, a library of random variable simulators is necessarily limited, and an analyst should be prepared to design customized functions. For example, the triangle density of Example 4.4 is not supported in the default `stats` package in R. We will show below how to simulate from this density. Apart from any practical questions, this area is quite interesting for its own sake, characterized by ingenuity and, sometimes, considerable mathematical depth. The more detailed survey of this topic in Ross (2019) is highly recommended.

The use a repeated samples (or *replications*) from some simulated random process is generally known as a *Monte Carlo method*. For example, suppose we wish to determine the expected value  $E[X]$  from some density  $f$ . It sometimes happens that this computation is analytically intractable, but a simulation from  $f$  is straightforward. In this case, if we simulate replicates  $X_1, \dots, X_N$ , then we have approximation

$$E[X] \approx \frac{\sum_{i=1}^N X_i}{N}.$$

The accuracy of this approximation can be assessed (this topic will be discussed in Chapter 12), and a value for  $N$  can be selected to yield a specific error tolerance (this topic will be discussed in Chapter 16).

The topic of simulation appropriately begins with the uniform distribution, and the linear congruential generator.

## 6.1 Linear Congruential Generators

Computer algorithms cannot produce truly random numbers (the very existence of randomness is an interesting question, but beyond the scope of this course). This is not necessarily a disadvantage. If our algorithm is deterministic, as it must be, then our results can be reproduced exactly by another analyst. In this case, it is perhaps more accurate to refer to a *pseudorandom* process.

Simulation often begins with the uniform distribution, which is typically generated by a *linear congruential generator*, which takes the form

$$x_{n+1} = (ax_n + b) \bmod P, \quad n = 0, 1, 2, \dots, \quad (6.1)$$

resulting in a sequence  $x_0, x_1, \dots$ . The numbers  $a$ ,  $b$  and  $P$  are fixed integers, and their informed choice is crucial. The number  $P$  is the *period*, and the initial value  $x_0$  is usually referred to as the *seed*. If  $a$  and  $b$  are both positive then the generator is *mixed congruential*, and if  $b = 0$  the generator is *multiplicative congruential*. Keeping track of the seed is crucial for reproducibility, since two linear congruential generators with the same parameters  $a, b, P$  will generate the same sequence using the same seed. Most computing environments, including R, allow the user to specify the seed, and allowing this option is good practice when designing randomized algorithms. The period  $P$  determines the range of the linear congruential generator, since any evaluation  $y = x \bmod P$  must be an integer between 0 and  $P - 1$  inclusive.

Suppose we set  $a = 2$ ,  $b = 3$ ,  $P = 20$ , with seed  $x_0 = 5$ . Then (6.1) produces the sequence

$$\begin{aligned} x_0 &= 5 \\ x_1 &= (2 \times x_0 + 3) \bmod 20 = 13 \\ x_2 &= (2 \times x_1 + 3) \bmod 20 = 9 \\ x_3 &= (2 \times x_2 + 3) \bmod 20 = 1 \\ x_4 &= (2 \times x_3 + 3) \bmod 20 = 5 \\ x_5 &= (2 \times x_4 + 3) \bmod 20 = 13 \\ x_6 &= (2 \times x_5 + 3) \bmod 20 = 9 \\ &\vdots \end{aligned}$$

yielding sequence 5, 13, 9, 1, 5, 13, 9, .... Note that the sequence returns to the seed value  $x_0 = x_4 = 5$  after 4 iterations of (6.1). As we would expect, the sequence begins to repeat itself, and will, in fact, repeat the sequence 5, 13, 9, 1 indefinitely. To what degree does the sequence depend on the seed? If we set  $x_0 = 19$ , we obtain sequence

$$19, 1, 5, 13, 9, 1, 5, 13, 9, \dots \quad (6.2)$$

so that the same four number sequence is repeated indefinitely. Interestingly, the sequence will never return to 19. On the other hand, if we set  $x_0 = 12$  we obtain sequence

$$12, 7, 17, 17, 17, \dots \quad (6.3)$$

so that 17 will be repeated indefinitely, and the sequence will never return to 12 or 7.

Since the period is  $P = 20$  we only need consider sequence elements  $0, 1, \dots, 19$ . If we set the seed in turn to each of these values, we find that when the seed is 2,7,12 or 17, the sequence eventually converges to 17, as in sequence (6.3). For all other seeds, the sequences will eventually cycle through 1,5,13,9 as in sequence (6.2).

To understand this behavior, it is important to note that the sequence is deterministic, in the sense that each possible value possesses exactly one other value that can follow it. For 17, that number happens to be 17 itself, that is, it is a *fixed point*, which by definition satisfies

$$x = (2 \times x + 3) \bmod 20.$$

We can then see that whenever the sequence enters the cycle 1,5,13,9 it will never leave, and whenever the sequence reaches 17, it will remain there indefinitely.

Clearly, the linear congruential generator possesses some interesting structure, but the type of behavior we have seen is clearly problematic, given its intended use. Ideally, we would like the sequence to avoid the type of behavior exemplified in (6.2) or (6.3), and consist of one single cycle of length  $P$ . This means that for any seed, the sequence will visit each possible value exactly once, and then return to the seed.

Fortunately, theory exists which informs the choice of parameters  $a, b, P$ . We have referred to  $P$  as the period. The attained period, as we seen, may be smaller. In the preceding example the attained period is either 4 or 1, depending on the seed. The largest possible attained period is called the *maximal period*, and we would like to have conditions under which the maximal period is attained. For linear congruential generators, this has been resolved with a great deal of precision and simplicity for the case  $P = 2^i$ . We present two results from Chambers (1977).

**Theorem 6.1** Suppose in (6.1)  $P = 2^i$ .

For the mixed congruential generator ( $a, b > 0$ ) the maximal period is  $P$ , and is attained if and only if

1.  $a \bmod 4 = 1$ ,
2.  $b$  is odd.

For the multiplicative congruential generator ( $a > 0, b = 0$ ) the maximal period is  $2^{i-2}$ , and is attained if and only if

1.  $a \bmod 8 = 3$  or  $5$ ,
2. the seed  $x_0$  is odd.

The multiplicative congruential generator may be more advantageous when the computing platform must be economical. ■

**Example 6.1** We examine two mixed congruential generators satisfying the conditions of Theorem 6.1. For both, we set  $P = 2^{20}$ , then consider separately  $(a, b) = (936001, 102295)$  and  $(5, 1)$  (we have  $936001 \bmod 4 = 1$ , for example).

The following R code produces plots with which to examine visually the degree of “randomness”. Figure 6.1 shows the first 100 elements of the sequence in various forms (top plots for  $(a, b) = (936001, 102295)$ , bottom plots for  $(a, b) = (5, 1)$ ). Roughly, the first generator appears generally random throughout. The second generator appears similarly unordered, except for short sequences of similar magnitude. This is due to the relatively small values of  $a$  and  $b$  compared to  $P$ .

```
> x = 1:100
> y = sin(x)
> plot(x,y)
>
> par(mfrow=c(2,2))
>
> P = 2^20
> a = 936001
> b = 102295
>
> x = rep(NA, 100)
>
> x0 = 1
> for (i in 1:100) {
+   x0 = (x0*a + b) %% P
+   x[i] = x0
+ }
> plot(1:100, x)
> plot(1:100, x, type='l')
>
> a = 5
> b = 1
> x0 = 1
> for (i in 1:100) {
+   x0 = (x0*a + b) %% P
+   x[i] = x0
+ }
```

```
> plot(1:100, x)
> plot(1:100, x, type='b')
```

The output is shown in Figure 6.1. ■

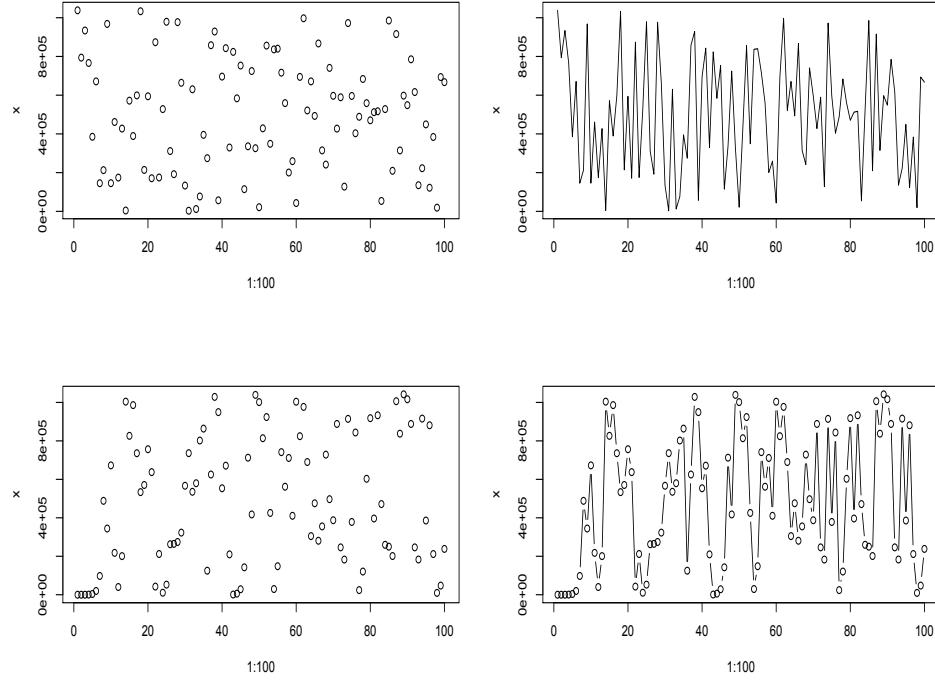


Figure 6.1: Sample sequences from linear congruential generators defined in Example 6.1. Top plots demonstrate  $(a, b) = (936001, 102295)$ , bottom plots demonstrate  $(a, b) = (5, 1)$ .

### 6.1.1 Uniform Random Number Generation

If we compute the first 10,000 variates from the generator

$$(a, b, P) = (936001, 102295, 2^{20})$$

of Example 6.1, the resulting histogram is shown in Figure 6.2. By Theorem 6.1 the maximal period  $2^{20} = 1,048,576$  is attained, and so the histogram accordingly covers approximately the range  $(0, 1048576)$ . In fact, the largest value observed in the sequence was 1,048,383, just below  $P$ . Simulated random variables with a uniform distribution on  $[0, 1]$  are given simply by dividing by  $P$ , that is

$$x_0/P, x_1/P, \dots, x_N/P$$

can be taken as a random sample from  $\text{unif}[0, 1]$ .

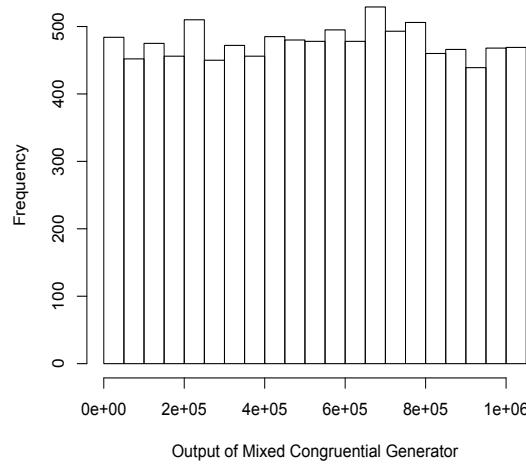


Figure 6.2: Histogram of the first 10,000 variates from the generator  $(a, b, P) = (936001, 102295, 2^{20})$  of Example 6.1.

## 6.2 The Inverse Transformation Method

Many simulation methods begin with uniform random variables. Suppose we are given a continuous CDF  $F$  and  $U \sim \text{unif}[0, 1]$ . Then  $F$  possesses inverse  $F^{-1}$ , since  $F$  is also increasing. This provides an elegant method of simulating a random variable with such a CDF.

**Theorem 6.2** Suppose  $F$  is a continuous CDF, and  $U \sim \text{unif}[0, 1]$ . Then the CDF of RV

$$X = F^{-1}(U) \tag{6.4}$$

is exactly  $F_X = F$ . ■

*Proof.* Under the hypothesis, the CDF of  $X$  is given by

$$X = F^{-1}(U). \tag{6.5}$$

The CDF of  $X$  is then

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= P(F^{-1}(U) \leq x) \\ &= P(U \leq F(x)) \\ &= F_U(F(x)). \end{aligned}$$

Recall that the CDF of  $U$  is  $F_U(u) = u$  for  $u \in [0, 1]$  (Section 4.11.1), so that completing the analysis gives

$$F_X(x) = F(x),$$

which completes the proof.  $\square$

A common application of this method is the simulation of exponentially distributed random variables.

**Example 6.2** If  $X \sim \exp(\lambda)$  then

$$F_X(x) = \begin{cases} 1 - \exp(-\lambda x) & ; \quad x \geq 0 \\ 0 & ; \quad x < 0 \end{cases} .$$

The inverse is obtained by solving

$$\begin{aligned} u &= 1 - \exp(-\lambda x), \\ \exp(-\lambda x) &= 1 - u, \\ x &= \frac{-\log(1 - u)}{\lambda}. \end{aligned}$$

So, given  $U \sim \text{unif}[0, 1]$ , if we set

$$X = F_X^{-1}(U) = \frac{-\log(1 - U)}{\lambda},$$

then  $X \sim \exp(\lambda)$ . Note that  $1 - U$  may be replaced by  $U$ , since both have the same distribution. ■

### 6.3 Simulation of Discrete Random Variables

The occurrence of an event with probability  $p$  is easily simulated with a random variable  $U \sim \text{unif}[0, 1]$ . For any interval  $I \subset [0, 1]$ , set

$$X = \begin{cases} 1 & ; \quad U \in I \\ 0 & ; \quad U \notin I \end{cases}$$

If  $I$  has length  $p$ , then  $X \sim \text{Bern}(p)$ .

The same idea may be used to simulate more complex discrete RVs. Suppose  $X$  has support  $\mathcal{S}_X = \{0, 1, \dots, M\}$ , and possesses PMF  $P_X(i) = p_i$ . Generate  $U \sim \text{unif}[0, 1]$ . If  $U \in [0, p_0)$  set  $X = 0$ . If  $U \in [p_0, p_0 + p_1)$  set  $X = 1$ . In general, set

$$X = \begin{cases} 0 & ; \quad U \in [0, p_0) \\ 1 & ; \quad U \in [p_0, p_0 + p_1) \\ \vdots & ; \quad \vdots \\ M & ; \quad U \in [p_0 + \dots + p_{M-1}, 1) \end{cases} .$$

This rule can be seen to be equivalent to

$$X = k \text{ if } U \in [F_X(k-1), F_X(k)] \tag{6.6}$$

and so is similar to the inverse transform method of the previous section.

**Example 6.3** Suppose  $X \sim \text{geom}(p)$ . Then  $F_X(k) = 1 - (1 - p)^k$ . Then using the method of Equation (6.6) we have

$$\begin{aligned} X &= \min\{k : U > 1 - (1 - p)^k\} \\ &= \min\{k : k > \frac{\log(1 - U)}{\log(1 - p)}\} \\ &= 1 + \lfloor \frac{\log(1 - U)}{\log(1 - p)} \rfloor, \end{aligned}$$

where  $\lfloor x \rfloor$  is the largest integer  $i \leq x$ , (referred to as the *floor function*). ■

## Chapter 7

# Stochastic Processes

A *stochastic process* may be defined as a (possibly uncountable) indexed collection of random variables  $\{X_t\}$ ,  $t \in \mathcal{T}$ . The index  $t$  usually represents time, although stochastic processes may also be used to describe random processes defined on some space.

Most stochastic processes are either *discrete time*, and take the form of a sequence  $X_1, X_2, \dots$ , or continuous time, and may be represented as a process  $X[t]$  on a subset  $t \in [0, \infty)$ , with  $X[t]$  being the value of the process at time  $t$ .

The *counting process* is a commonly encountered continuous time stochastic process

**Definition 7.1** A *counting process* is a stochastic process  $N(t)$  defined on  $t \in [0, \infty)$  satisfying (i)  $N(0) = 0$ ;  $N(t) \in \mathbb{I}$ ;  $N(t)$  is nondecreasing in  $t$ . Let  $T_0 = 0$  and  $T_i = \inf\{t : N(t) \geq i\}$ ,  $i \geq 1$ . A counting process is a *renewal process* if the differences  $T_i - T_{i-1}$ ,  $i \geq 1$  are independent and identically distributed. ■

Usually,  $N(t)$  represents the number of events in a sequence which have occurred by time  $t$ . It is helpful to think of  $N(t)$  as an *arrival process*, as though we were marking the arrival of customers at a queue starting at  $N(0) = 0$  at time  $t = 0$ . Then for  $t > s$   $N(t) - N(s)$  is the number of arrivals in time interval  $(s, t]$ . Note that  $N(t)$  contains perfect information regarding the times at which events occur.

## 7.1 Poisson Process

One of the most important stochastic process models is the *Poisson process*, which relies on the following definitions. A counting process  $N(t)$  has *independent increments* if the quantities  $N(t_1) - N(s_1)$  and  $N(t_2) - N(s_2)$  are independent whenever  $s_1 < t_1 < s_2 < t_2$ . In addition,  $N(t)$  is *stationary* (or has *stationary increments*) if the distribution of  $N(t) - N(s)$  depends only on  $t - s$  for any  $s < t$ .

We next define a *Poisson process*.

**Definition 7.2 (Poisson Process (Definition 1))** A counting process  $N(t)$  is a (homogenous) Poisson process with rate  $\lambda$  if the following conditions hold:

- (i)  $N(0) = 0$ ,
- (ii)  $N(t)$  has independent and stationary increments,
- (iii)  $P(N(s) = 1) = \lambda s + o(s)$ ,
- (iv)  $P(N(s) > 1) = o(s)$ .

■

The model can be generalized by allowing rate  $\lambda$  to vary in time, in which case the Poisson process is inhomogeneous. Usually, the process is assumed to be homogeneous unless explicitly stated otherwise. The properties (ii)–(iii) essentially state that the expected number of arrivals between times  $t$  and  $t + s$  is approximately  $\lambda s$ , so that  $\lambda$  becomes the arrival rate. Definition 7.2 is a strong one, since it may be shown that it implies very specific distributional properties. In fact, it may be shown that Definition 7.2 is equivalent to the following alternative definition of a Poisson process: holds we must conclude that  $N(t+s) - N(t)$  has a Poisson distribution with mean  $\lambda s$  (see Ross (1996), Chapter 2).

**Definition 7.3 (Poisson Process (Definition 2))** A counting process  $N(t)$  is a (homogenous) Poisson process with rate  $\lambda$  if the following conditions hold:

- (i)  $N(0) = 0$ ,
- (ii)  $N(t)$  has independent and stationary increments,
- (iii) The number of events in any interval of length  $s$ , specifically  $N(t+s) - N(t)$ , has a Poisson distribution with mean  $\lambda s$ .

■

Then let  $X_1, X_2, \dots$  be the interarrival times of a Poisson process  $N(t)$ . The distribution of  $X_1$  is easily determined by noting that

$$P(X_1 > s) = P(N(s) - N(0) = 0) = \exp(-\lambda s),$$

making use of Definition 7.3 (iii), so that  $X_1$  has an exponential distribution with mean  $1/\lambda$ . We may then write

$$\begin{aligned} P(X_2 > s \mid X_1 = t) &= P(\text{no arrivals in } (t, t+s] \mid X_1 = t) \\ &= P(\text{no arrivals in } (t, t+s]) \\ &= P(N(t+s) - N(t) = 0) \\ &= \exp(-\lambda s). \end{aligned}$$

Here, we make use Definition 7.3 (ii) to argue that the arrivals on interval  $(t, t+s]$  and the arrivals on interval  $[0, t]$  are independent, so that  $P(\text{no arrivals in } (t, t+s] \mid X_1 = t) = P(\text{no arrivals in } (t, t+s])$ . Then  $P(N(t+s) - N(t) = 0) = \exp(-\lambda s)$  follows directly from Definition 7.3 (iii). This means that  $X_2$  also has an exponential distribution with mean

$1/\lambda$ , and is independent of  $X_1$ . If we continue the argument, we conclude that the interarrival times are *iid* observations from exponential distribution with mean  $1/\lambda$ .

The Poisson process is of importance in stochastic modelling for much the same reason that the normal distribution is of importance in modeling noise processes. The latter is the limit of a superposition of an arbitrarily large number of independent (or weakly dependent) noise processes. Similarly, the Poisson process is a limit of a superposition of an arbitrarily large number of arrival processes. This means that an arrival process which is really an aggregation of a large number of essentially separate arrival processes of general type will resemble a Poisson process. Derivations of this fact can be found in Section 5.9 of Karlin and Taylor (1975) or Section XI.3 of Feller (1971).

**Example 7.1** Customers arrive in a shop according to a Poisson process with rate  $\lambda = 0.35/\text{minute}$ . The probability that no customers arrive within a specific 5 minute interval is equal to  $P(Y = 0)$ , where  $Y \sim \text{Pois}(L)$ , and  $L = 5 \times 0.35 = 1.75$  is the expected number of arrivals within 5 minutes. This gives  $P(Y = 0) = e^{-1.75} \approx 0.174$ . ■

## 7.2 Markov Chains

The *Markov chain* is a discrete time stochastic process. The defining property is very much related to the memoryless property of geometric or exponential random variables (Sections 4.10.4 and 4.11.2). It is commonly referred to as well as the memoryless property, also the *Markovian property*,

**Definition 7.4** Suppose we are given a discrete time stochastic process  $X_n \in \mathcal{X}$ ,  $i = 0, 1, 2, \dots$ , which assumes values in a discrete *state space*  $\mathcal{X}$ . Without loss of generality we have either a finite state space  $\mathcal{X} = \{0, 1, \dots, n\}$  or countable state space  $\mathcal{X} = \{0, 1, \dots\}$ . Then  $X_i$  is a *Markov chain* if the following *memoryless property* holds:

$$\begin{aligned} P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \dots, X_1 = i_1, X_0 = i_0) \\ = P(X_{n+1} = j | X_n = i) \\ = P_{ij}. \end{aligned}$$

The quantity  $P_{ij}$  is called the *transition probability* from state  $i$  to state  $j$ . We also have *transition probability matrix* (or *transition matrix*)

$$P = \begin{bmatrix} P_{00} & P_{01} & P_{02} & \cdots \\ P_{10} & P_{11} & P_{12} & \cdots \\ \vdots & \vdots & \vdots & \\ P_{i0} & P_{i1} & P_{i2} & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix}$$

Row  $i$  of transition matrix  $P$  is equivalent to the conditional probability

$$P(X_{n+1} = j \mid X_n = i) = P_{ij}, \quad j \in \mathcal{X}.$$

Note also that  $P$  will be a matrix of infinite dimension when  $\mathcal{X}$  is countable. We also have no difficulty conceiving of  $P$  as ‘doubly infinite’ when the state space is the set of positive and negative integers  $\{\dots, -2, -1, 0, 1, 2, \dots\}$ , which requires no important change of Definition 7.4.

**Example 7.2** We start with an example of a two-state Markov chain, which, despite its simplicity, demonstrates a number of important features of Markov chains. Formally, we have state space  $\mathcal{X} = \{0, 1\}$ . However, we lose nothing by replacing the notation of Definition 7.4 with something more intuitive.

For example, the time index  $i = 0, 1, 2, \dots$  may represent a sequence of days, and we may wish to define a simple infection model, in which state  $i = 0$  represents a *healthy state H* and  $i = 1$  represents a *sick state I* (due to, say, an infection). The transition matrix is therefore the  $2 \times 2$  matrix

$$P = \begin{bmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{bmatrix}.$$

However, the true degrees of freedom of  $P$  is 2, since each row is constrained, as a probability distribution, to sum to 1 (such a matrix is known as a *stochastic matrix*). We can therefore write, without loss of generality,

$$P = \begin{bmatrix} 1 - \alpha & \alpha \\ \beta & 1 - \beta \end{bmatrix}. \quad (7.1)$$

for two numbers  $\alpha, \beta \in [0, 1]$ . This means that if a subject is healthy on day  $i$ , he/she is sick on day  $i + 1$  with probability  $\alpha$ , and if the subject is sick on day  $i$ , he/she is sick on healthy in day  $i + 1$  with probability  $\beta$ . The state transition diagram for infection model is shown in Figure 7.1.

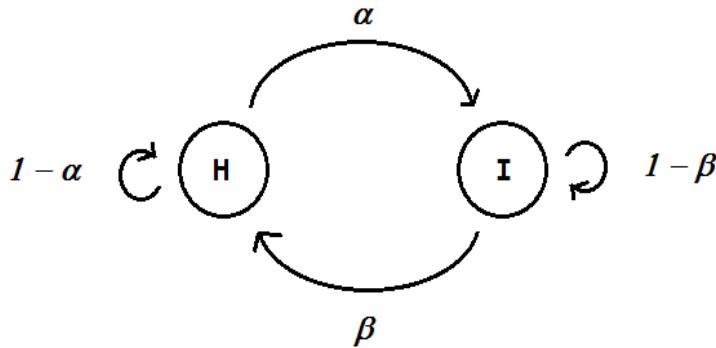


Figure 7.1: State transition diagram for infection model of Example 7.2.

Is this a reasonable model? First, we note that when the subject enters state  $H$ , he/she remains there for a geometrically distributed waiting time, with mean  $\alpha^{-1}$  (see Section 4.10.4). If we suppose that acquiring an infection is a consequence of a chance exposure, which happens on any given day with probability  $\alpha$ , then the memoryless ‘coin toss’ model for the waiting time until infection would be reasonable.

On the other hand, the infection lifetime also follows a geometric distribution, but with mean  $\beta^{-1}$ . Presumably, clinical experience would guide the choice of  $\beta$ , setting

$$\beta^{-1} = E[\text{infection lifetime}].$$

However, whether the geometric distribution adequately models an infection lifetime would be an important question to resolve. ■

The transition probability  $P_{ij}$  may be more formally referred to as the *one-step transition probability*, since it describes transition following a single time step. We may also describe the  $k$ -step transition probability

$$P(X_{n+k} = j \mid X_n = i) = P_{ij}^k,$$

noting that this probability does not depend on  $n$ . We will demonstrate this computation for  $k = 2$ . Recall the *law of total probability* of Section 2.7. Set

$$E = \{X_{n+2} = j\}, \quad B = \{X_n = i\},$$

in Equation (2.2), then define partition

$$A_k = \{X_n = k\} \text{ for all } k \in \mathcal{X}.$$

This gives

$$\begin{aligned} P_{ij}^2 &= P(X_{n+2} = j \mid X_n = i) \\ &= \sum_{k \in \mathcal{X}} P(X_{n+2} = j \mid X_{n+1} = k, X_n = i) P(X_{n+1} = k \mid X_n = i). \end{aligned} \quad (7.2)$$

Then consider each term in the summation. Recall by the Markovian property of Definition 7.4 that the distribution of  $X_{n+2}$  given history

$$X_{n+1}, X_n, \dots, X_1, X_0$$

depends only on the most recent state  $X_{n+1}$ . We therefore have

$$P(X_{n+2} = j \mid X_{n+1} = k, X_n = i) = P(X_{n+2} = j \mid X_{n+1} = k) = P_{kj}. \quad (7.3)$$

The remaining quantity is simply the one step transition probability

$$P(X_{n+1} = k \mid X_n = i) = P_{ik}. \quad (7.4)$$

Substituting (7.3) and (7.4) into (7.2) yields

$$P_{ij}^2 = \sum_{k \in \mathcal{X}} P_{ik} P_{kj}. \quad (7.5)$$

This is a particularly important relationship, since we can recognize (7.5) as the result of matrix multiplication. We summarize this in the following definition.

**Definition 7.5** The *k-step transition probability* from state  $i$  to  $j$  is the probability that a Markov chain in state  $i$  occupies state  $j$  after exactly  $k$  transitions. Formally,

$$P(X_{n+k} = j \mid X_n = i) = P_{ij}^k,$$

The  $k$ -step transition probability matrix  $P^k$  has value  $P_{ij}^k$  at (row,column)  $(i, j)$ , and can be calculated by iteratively multiplying  $P$   $k$  times:

$$P^k = [P]^k$$

■

### 7.2.1 Maze Example

There has always been considerable interest in the skill with which rats navigate mazes. Markov chains provide an analytical tool with which the translate experimental observations into relevant conclusions.

For example, the term ‘memoryless’ implies an inability to learn. A Markov chain should be able to model navigational behavior which is more or less random, implying an inability to learn by trial and error. Therefore, we should be able to decide whether or not experimental data is explainable by truly memoryless behavior.

Consider the maze shown in Figure 7.2. A rat is introduced to the maze at the ‘Start’ label, and is provided some incentive to reach the ‘Finish’ label. There are 4 nodes of decision, labeled  $a, b, c, d$ , and a terminal node  $e$ . Node  $a$  is the first node encountered by the rat.

How should we model the rat’s navigation? It seems reasonable to define the state space  $\mathcal{X} = \{a, b, c, d, e\}$ , modeling transitions between decision points. We might suppose that if the rat’s behavior is truly random, whenever it reaches a decision node, it simply chooses any available direction with equal probability, and otherwise it walks along its chosen path.

Assuming the entrance is closed after entry, from node  $a$  the rat can proceed N(orth), E or S. If it proceeds N, it next reaches node  $b$ , otherwise it must return to  $a$ . This gives transition probabilities

$$P_{aa} = 2/3, \quad P_{ab} = 1/3, \quad P_{ak} = 0 \text{ for } k = c, d, e.$$

From node  $b$  the rat can proceed W, E or S. If it proceeds W it returns to  $a$ , if it proceeds E it reaches a dead end, and so must return to  $b$ , and if it proceeds S it reaches  $c$ . This gives transition probabilities

$$P_{ba} = 1/3, \quad P_{bb} = 1/3, \quad P_{bc} = 1/3, \quad P_{bk} = 0 \text{ for } k = d, e.$$

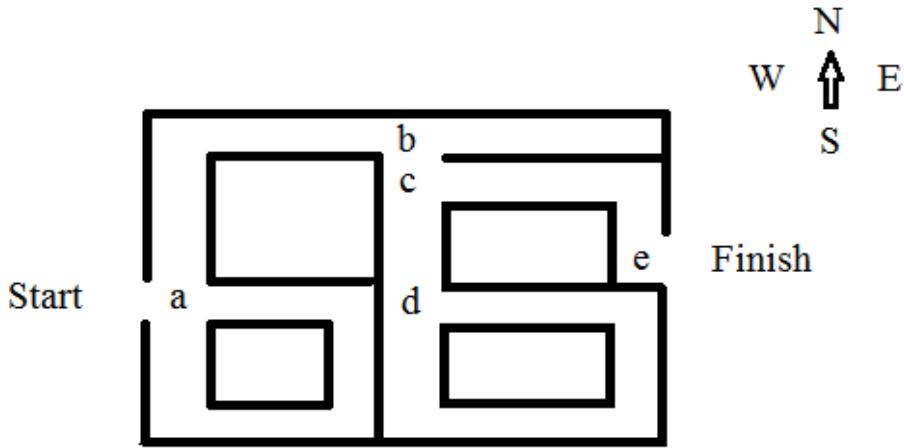


Figure 7.2: Maze diagram for Section 7.2.1.

Continuing in this way, and ordering states  $a, b, c, d, e$  as  $0, 1, 2, 3, 4$  we have transition matrix

$$P = \begin{bmatrix} & a & b & c & d & e \\ a & 2/3 & 1/3 & 0 & 0 & 0 \\ b & 1/3 & 1/3 & 1/3 & 0 & 0 \\ c & 0 & 1/3 & 0 & 1/3 & 1/3 \\ d & 0 & 0 & 1/3 & 1/3 & 1/3 \\ e & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (7.6)$$

Note that  $P_{ee} = 1$ , so that once the Markov chain reaches state  $e$  it remains there indefinitely, which is commonly interpreted as a termination of the process. Such a state is referred to as an *absorbing state*.

After some reflection, we might ask if a rat, when reaching a decision node, is likely to respond by reversing direction. We have assumed that the rat is as likely to do this as to choose any of the remaining directions.

Let's consider what happens if we try to rule out this behavior. If the rat reaches node  $a$  while traveling N, if we rule out direction reversals, the rat will either proceed N, reaching  $b$ , or proceed E, returning to  $a$ . The nonzero transition probabilities are now  $P_{aa} = 1/2$ ,  $P_{ab} = 1/2$ , which differ from those given in (7.6). Unfortunately, we can see that these transition probabilities depend on the direction of travel. If the rat reaches node  $a$  by traveling S, barring direction reversals, the rat can only proceed S or E, forcing a return to  $a$ , suggesting that  $P_{aa} = 1$ . The problem is that the Markovian property is being violated. To say that the transition probabilities depend on the direction of travel is equivalent to saying that they depend not only on the current state (node), but also on the previous state (node). After all, if the rat reaches node  $a$  traveling S, it must previously have been at node  $b$ , whereas if it reaches node  $a$  traveling N, it must previously have been at node  $a$ . Definition 7.4 is therefore not satisfied.

Fortunately, it is possible to rule out direction reversals while maintaining the memoryless property. This can be done by expanding the state space (sometimes referred to as *Markovianizing* a process). In particular we include in the definition of the state both the node and the direction of approach. That is, the rat transitions to state  $a - N$  by arriving at node  $a$  traveling N. At state  $a - N$  the rat proceeds N or E with equal probability. By proceeding N the rat reaches state  $b - E$ , and by proceeding E the rat returns to state  $a - N$ . This gives nonzero transition probabilities

$$P_{a-N,a-N} = 1/2, \quad P_{a-N,b-E} = 1/2.$$

Node  $a$  must be expanded into states  $a - E$ ,  $a - N$ ,  $a - S$  and  $a - W$ . Our protocol is to designate  $X_0 = a - E$  as the initial state, although the definition of the Markov chain model remains the same for any initial state. Once the rat leaves state  $a - E$  it does not return.

Similarly, state  $b$  is expanded into state  $b - E$ ,  $b - W$  and  $b - N$ . Since the rat cannot approach  $b$  traveling S, no state  $b - S$  is needed. From state  $b - E$  the rat can proceed E or S. If the rat proceeds E, it meets a dead end, and must return to node  $b$  traveling W, thus transitioning to state  $b - W$ . Otherwise, it arrives at node  $c$  traveling S, thus transitioning to node  $c - S$ . This gives nonzero transition probabilities

$$P_{b-E,b-W} = 1/2, \quad P_{b-E,c-S} = 1/2.$$

Continuing in this way, we may deduce the transition probabilities given in Table 7.1.

Table 7.1: Transition probabilities for Markov chain model of Section 7.2.1

$P_{ij} =$	a-E	a-N	a-S	a-W	b-N	b-E	b-W	c-N	c-S	d-N	d-S	d-W	e-S
a-E	0	1/3	0	1/3	0	1/3	0	0	0	0	0	0	0
a-N	0	1/2	0	0	0	1/2	0	0	0	0	0	0	0
a-S	0	1/2	0	1/2	0	0	0	0	0	0	0	0	0
a-W	0	0	0	1/2	0	1/2	0	0	0	0	0	0	0
b-N	0	0	1/2	0	0	0	1/2	0	0	0	0	0	0
b-E	0	0	0	0	0	0	1/2	0	1/2	0	0	0	0
b-W	0	0	1/2	0	0	0	0	0	1/2	0	0	0	0
c-N	0	0	0	0	1/2	0	0	0	0	0	0	0	1/2
c-S	0	0	0	0	0	0	0	0	0	0	1/2	0	1/2
d-N	0	0	0	0	0	0	0	1/2	0	1/2	0	0	0
d-S	0	0	0	0	0	0	0	0	0	1/2	0	1/2	0
d-W	0	0	0	0	0	0	0	1/2	0	0	0	1/2	0
e-S	0	0	0	0	0	0	0	0	0	0	0	0	1

A Markov chain is easily simulated, using the methods of Chapter 6. An example of a single path is shown in Figure 7.3. In this example there are 26 transitions. A total of 27 states are visited, including initial and final states  $a - E$  and  $e - S$ . We note some 'cycling' behavior around node  $a$  at transitions 1 and 13, and around node  $d$  at transitions 8 and 21. This is characteristic of memoryless behavior.

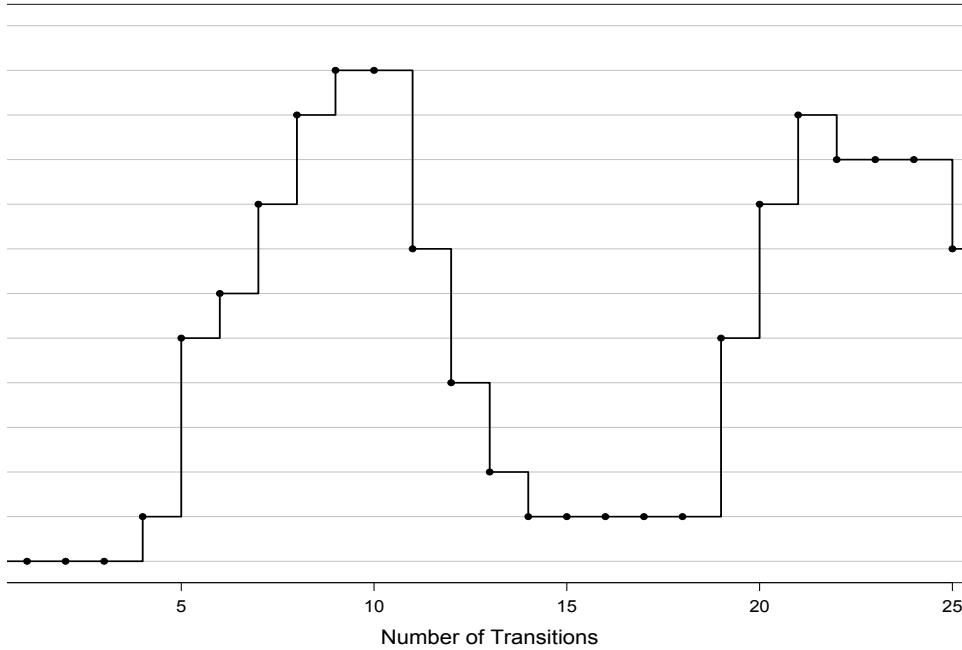


Figure 7.3: Simulated Markov chain navigation for maze example of Section 7.2.1, based on transition probabilities of Table 7.1.

### 7.2.2 Distributional Properties of Markov Chains

All distributional properties of a Markov chain follow from transition matrix  $P$ . In principle, these properties may be studied by simulating *sample paths*, using some large number of replications  $N$ . For example, in the maze example suppose  $T$  is the number of transitions required to reach terminal state  $e - S$  from initial state  $a - E$ . This is a random survival time, and examining Figure 7.3 we can see that the minimum possible value is  $T = 3$ , attained exclusively by path

$$a - E \rightarrow b - E \rightarrow c - S \rightarrow e - S. \quad (7.7)$$

Note that  $T$  is one less than the number of states in a path. Suppose we observe a rat solve the maze with  $T = 3$ . If the rat has already had some experience with the maze, we may suspect that it has ‘learned’ the shortest route. On the other hand,  $T = 3$  is an outcome which may occur under the memoryless model defined in Table 7.1. We can form some judgement by calculating the probability  $P(T = 3)$ . If it is very small, we conclude that the rat is unlikely to have achieved the optimal path by chance.

Simulations may be used to estimate  $P(T = 3)$ . If out of  $N$  simulated

paths,  $X$  equal the optimal path, then

$$P(T = 3) \approx X/N.$$

However, the Markov chain model proves to be a very powerful analytic tool, and many interesting properties may be calculated exactly. When this is possible, it is to be preferred to simulation.

For example, the probability of a path is easily calculated.

**Theorem 7.1** Let

$$i_0 \rightarrow i_1 \dots \rightarrow i_{m-1} \rightarrow i_m$$

be any path of  $m$  transitions. The probability of the path, given initial state  $X_0 = i_0$  is

$$\begin{aligned} P(X_m = i_m, X_{m-1} = i_{m-1}, \dots, X_1 = i_1 \mid X_0 = i_0) \\ = P_{i_0, i_1} \times P_{i_1, i_2} \times \dots \times P_{i_{m-2}, i_{m-1}} \times P_{i_{m-1}, i_m} \end{aligned} \quad (7.8)$$

for any such path. ■

*Proof.* . We may write

$$\begin{aligned} P(X_m = i_m, X_{m-1} = i_{m-1}, \dots, X_1 = i_1, X_0 = i_0) \\ = P(X_m = i_m \mid X_{m-1} = i_{m-1}, \dots, X_1 = i_1, X_0 = i_0) \\ \times P(X_{m-1} = i_{m-1}, \dots, X_1 = i_1, X_0 = i_0). \end{aligned} \quad (7.9)$$

By the Markov property

$$\begin{aligned} P(X_m = i_m \mid X_{m-1} = i_{m-1}, \dots, X_1 = i_1, X_0 = i_0) \\ = P(X_m = i_m \mid X_{m-1} = i_{m-1}) \\ = P_{i_{m-1}, i_m}. \end{aligned} \quad (7.10)$$

Substituting (7.10) into (7.9) yields

$$\begin{aligned} P(X_m = i_m, X_{m-1} = i_{m-1}, \dots, X_1 = i_1, X_0 = i_0) \\ = P_{i_{m-1}, i_m} P(X_{m-1} = i_{m-1}, \dots, X_1 = i_1, X_0 = i_0). \end{aligned} \quad (7.11)$$

We may apply essentially the same argument to the quantity  $P(X_{m-1} = i_{m-1}, \dots, X_1 = i_1, X_0 = i_0)$  in (7.11), giving

$$\begin{aligned} P(X_m = i_m, X_{m-1} = i_{m-1}, \dots, X_1 = i_1, X_0 = i_0) \\ = P_{i_{m-2}, i_{m-1}} P_{i_{m-1}, i_m} P(X_{m-2} = i_{m-2}, \dots, X_1 = i_1, X_0 = i_0). \end{aligned}$$

Repeating enough times, we have

$$\begin{aligned} P(X_m = i_m, X_{m-1} = i_{m-1}, \dots, X_1 = i_1, X_0 = i_0) \\ = P_{i_0, i_1} \times P_{i_1, i_2} \times \dots \times P_{i_{m-2}, i_{m-1}} \times P_{i_{m-1}, i_m} \times P(X_0 = i_0). \end{aligned} \quad (7.12)$$

We condition both sides of (7.12) on  $\{X_0 = i_0\}$  by dividing by  $P(X_0 = i_0)$ , which yields (7.8). □

Note that the probability of a path starting at any time index may be calculated using the method of Theorem 7.1.

**Example 7.3** For the Markov chain of Table 7.1, if  $T$  is the number of transitions required to reach terminal state  $e - S$ , then, as discussed earlier in this section, the minimum value  $T = 3$  is reached exclusively by the path of (7.7). By Theorem 7.1, we have

$$\begin{aligned} P(T = 3) &= P(X_3 = e - S, X_2 = c - S, X_1 = b - E \mid X_0 = a - E) \\ &= P_{a-E,b-E} P_{b-E,c-S} P_{c-S,e-S} \\ &= 1/3 \times 1/2 \times 1/2 \\ &= 1/12. \end{aligned}$$

Therefore, under the memoryless model, the optimal path is attained with probability  $1/12$ , and would therefore be consistent with the memoryless model. ■

It turns out that the exact distribution of  $T$  is readily obtained. Recall the  $k$ -step transition probability

$$P_{a-E,e-S}^k = P(X_k = e - S \mid X_0 = a - E).$$

Clearly, if  $X_k = e - S$ , we must have  $T \leq k$  (the path may have entered state  $e - S$  before the  $k$ th transition). However, recall that  $e - S$  is an absorbing state, so that if  $X_j = e - S$  for some  $j$ , we must also have if  $X_k = e - S$  for all  $k \geq j$ . This means that if  $T \leq k$  we must also have  $X_k = e - S$ , so that

$$\{T \leq k\} = \{X_k = e - S\}$$

and so we have CDF

$$F_T(k) = P(T \leq k) = P(X_k = e - S) = P_{a-E,e-S}^k.$$

Note that we are implicitly conditioning the distribution of  $T$  on the event  $\{X_0 = a - E\}$ . Then, the  $k$ -step probabilities  $P_{a-E,e-S}^k$  can be obtained by matrix multiplication of  $P$ . Interpreting Table 7.1 as a  $13 \times 13$  matrix  $P$ , we have

$$F_T(k) = P^k[1, 13],$$

from which we get the PMF

$$p_T(k) = F_T(k) - F_T(k-1) = P^k[1, 13] - P^{k-1}[1, 13],$$

where the matrix  $P^0$  is taken to be the identity matrix. Figure 7.4 shows the values of  $P(T = k)$ . We may compute the expected value

$$E[T] \approx 11.33.$$

On average, under the memoryless model, a rat requires 11.33 transitions to solve the maze. However, note that the tail of the distribution extends well above this value. For example, we have the probability  $P(T \geq 25) \approx 0.083$  and  $P(T \geq 35) = 0.027$ , so observing a solution time well above the mean will not be a rare occurrence.

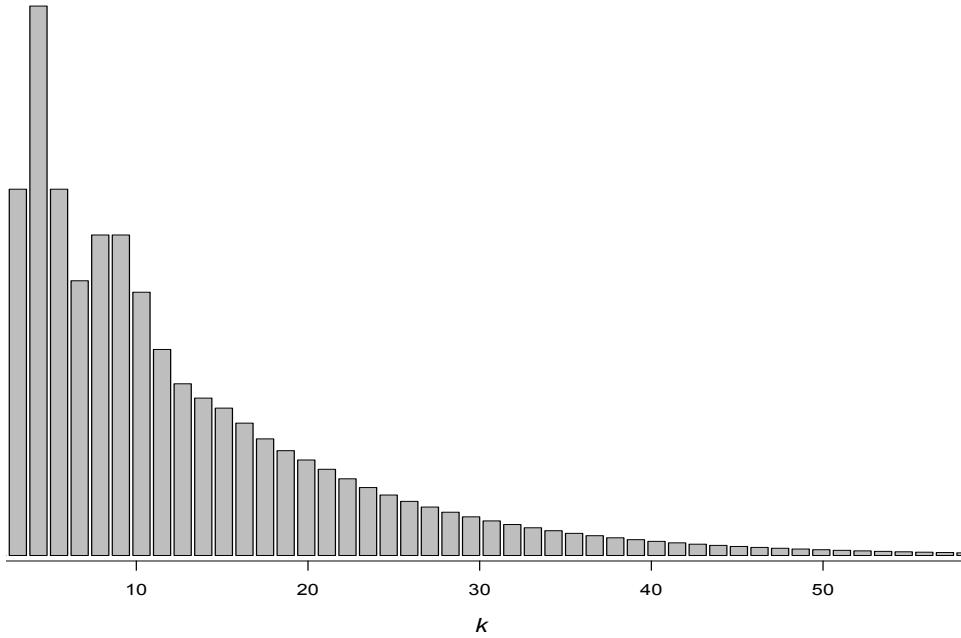


Figure 7.4: PMF  $p_T(k) = P(T = k)$  of  $T$ , the number of transitions required to reach terminal state  $e - S$ , for maze example of Section 7.2.1, based on transition probabilities of Table 7.1.

### 7.2.3 Balance Equations and Steady States

It is of some interest to contrast the *long run* behavior of the two-state Markov chain of Example 7.2 with the maze example of Section 7.2.1. The first model can be expected to fluctuate between *Healthy* and *Infected* states indefinitely, while the maze model has a stopping condition. In the former case, we may be interested in knowing the long run proportion of time spent in each state.

Suppose we have counting process

$$N_j(k) = \text{The number of transitions into state } j \text{ after the } k\text{th transition.}$$

A long run frequency (formally, a *steady state frequency*) would then be defined by

$$\pi_j = \lim_{k \rightarrow \infty} \frac{N_j(k)}{k}. \quad (7.13)$$

See Section B.3 for a formal definition of limits. A number of mathematical questions lurk here. Does the limit always exist? If so, under what conditions is  $\pi_j$  zero or positive? Last, and far from least, does this quantity depend on the initial state? To formally resolve these questions requires some amount

of mathematical theory, even when dealing with relatively simple models. While this would be beyond the scope of this course, we can give some insight into the issues.

For the maze model, the long run frequencies  $\pi_j$  are easy to deduce, in particular,  $\pi_{e-S} = 1$ , and  $\pi_k = 0$  for all other states. Deducing  $\pi_j$  for general models requires acknowledgement of an apparently obvious fact, that is,

$$\begin{aligned} & \text{Number of times a Markov chain enters a state} \\ &= \text{Number of times a Markov chain exits a state}, \end{aligned} \quad (7.14)$$

(plus zero or one). The value of this statement becomes more apparent if we think in terms of *rates*. Clearly,  $\pi_j$  can be interpreted as the *occupancy rate* of state  $j$ , but also as its *entrance rate* and the *exit rate*. Next, we may consider the rate at which the Markov chain transitions from states  $i$  to  $j$ . There are two components to this. First, to transition from  $i$  to  $j$ , the Markov chain must first enter  $i$ . This occurs at rate  $\pi_i$ . Second, given that the Markov chain is in  $i$ , it transitions from  $i$  to  $j$  with probability  $P_{ij}$ . The rate of transition from  $i$  to  $j$  is therefore  $\pi_i P_{ij}$ .

We are now in a position to use (7.14). We recognize the exit rate for state  $j$  as  $\pi_j$ . The entrance rate, on the other hand, can be given as the sum of all other transition rates *into* state  $j$ , that is,  $\pi_i P_{ij}$  for  $i \in \mathcal{X}$  (this includes  $i = j$ , when  $P_{jj} > 0$ ). Equation (7.14) then yields the *balance equation*

$$\pi_j = \sum_{i \in \mathcal{X}} \pi_i P_{ij}. \quad (7.15)$$

**Example 7.4** Consider the two-state Markov chain of Example 7.2. We write a balance equation for each state, yielding

$$\begin{aligned} \pi_0 &= \pi_0 P_{00} + \pi_1 P_{10} \\ \pi_1 &= \pi_0 P_{01} + \pi_1 P_{11}, \end{aligned}$$

which, after substituting transition probabilities (7.1) gives

$$\begin{aligned} \pi_0 &= \pi_0(1 - \alpha) + \pi_1\beta \\ \pi_1 &= \pi_0\alpha + \pi_1(1 - \beta). \end{aligned}$$

Rewriting the first equation yields

$$\frac{\pi_0}{\pi_1} = \frac{\beta}{\alpha}. \quad (7.16)$$

Since we must have  $\sum_i \pi_i = 1$ , only one balance equation is actually needed to solve for  $(\pi_0, \pi_1)$ , and we obtain

$$\pi_0 = \frac{\beta}{\alpha + \beta}, \quad \pi_1 = \frac{\alpha}{\alpha + \beta}.$$

That the frequencies  $\pi_0, \pi_1$  should possess ratio  $\beta/\alpha$  is to be expected. The time spent in each state prior to transition is geometrically distributed with means  $1/\alpha$  and  $1/\beta$  respectively. The ratio  $\pi_0/\pi_1$  should then be the ratio of the means, which is confirmed by (7.16). ■

### 7.3 Birth and Death Processes

We start with the following definition:

**Definition 7.6** A stochastic process  $X(t)$  on continuous time  $t \in [0, \infty)$  is a *continuous-time Markov chain* if  $X(t) \in \mathcal{S}$  for some discrete state space  $\mathcal{X}$  and

$$\begin{aligned} P(X(t+s) = j \mid X(s) = i, X(u) = x(u), u \in [0, s)) \\ = P(X(t+s) = j \mid X(s) = i) \end{aligned}$$

for  $i, j \in \mathcal{X}$ . ■

The theory of the continuous-time Markov chain is considerably simplified by the following Theorem:

**Theorem 7.2** The following properties hold for the exponential distribution:

- (i) The exponentially distributed random variable is the unique memoryless waiting time with support  $[0, \infty)$ .
- (ii) If  $X_1, \dots, X_n$  are independent exponentially distributed random variables with rates  $\lambda_1, \dots, \lambda_n$  then  $Y = \min_i X_i$  is exponentially distributed with rate  $\sum_i \lambda_i$ .
- (iii) In addition,  $P(Y = X_j) = \lambda_j / \sum_i \lambda_i$ . ■

*Proof.* . (i) A waiting time  $X$  on  $[0, \infty)$  is memoryless if and only if  $P(X > t + s \mid X > t) = P(X > s)$ , or  $P(X > t + s) = P(X > s)P(X > t)$ , for all  $s, t \geq 0$ . That this property is satisfied by the exponential distribution is easily verified by substituting  $\bar{F}(x) = \exp(-\lambda x)$ . To prove the converse, suppose a memoryless waiting time on support  $[0, \infty)$  has distribution function  $F$ . Letting  $S(u) = \log(\bar{F}(u))$ , the memoryless property implies  $S(t+s) = S(t) + S(s)$ . Since  $S$  is monotone, a solution to this equation must be of the form  $S(t) = ct$ , which completes the proof.

(ii)  $P(Y > t) = P(\cap_i \{X_i > t\}) = \prod_i P(X_i > t) = \prod_i \exp(-\lambda_i t) = \exp(-t \sum_i \lambda_i)$ .

(iii) For  $n = 2$ , we may evaluate  $P(Y = X_1) = P(X_1 < X_2) = \int_{x_1 < x_2} f(x_1 \mid \lambda_1) f(x_2 \mid \lambda_2) dx_1 dx_2 = \lambda_1 / (\lambda_1 + \lambda_2)$ . The extension to  $n \geq 2$  follows from (ii). □

The continuous-time Markov chain is perhaps best introduced by the special case of the *birth and death process*, which satisfies Definition 7.6, with the additional requirements that  $\mathcal{X} = \{0, 1, 2, \dots\}$  and that state transitions can occur only between adjacent integers.

Suppose the process is at state  $i$  at time  $t$ , that is,  $X(t) = i$ . We can imagine two competing *waiting times*  $W_+$  and  $W_-$ . If  $W_+ < W_-$  then the

process moves to state  $i+1$  after waiting time  $W_+$ , that is  $X(t+W_+) = i+1$ . Similarly, if  $W_- < W_+$  then the process moves to state  $i - 1$  after waiting time  $W_-$ , that is  $X(t + W_-) = i - 1$ . At state  $i = 0$  there is only one waiting time  $W_+$ , but the process is otherwise the same. In order for the memoryless property of Definition 7.6 to hold, we need only assume that these waiting times are exponential. Once we specify exponential rates  $\lambda_j$ ,  $\mu_j$  for the distributions of  $W_+$  and  $W_-$  for each state  $j$ , the process is completely defined. Accordingly, we call the parameters  $\lambda_j$  *birth* or *arrival* rates, and  $\mu_j$  are the *death* or *departure* rates.

As for the Markov chain, we may construct balance equations. Suppose  $P_j$  is the long run proportion of time spent in state  $j$ . Then  $\lambda_j$  and  $\mu_j$  are the rates of transition from  $j$  to  $j+1$  and  $j-1$  respectively. We may accommodate the state  $j = 0$  by assuming  $\mu_0 = 0$ . After equating entrance and exit rates in the same manner as for Equation (7.15) the balance equations become

$$\begin{aligned}\lambda_0 P_0 &= \mu_1 P_1, \\ (\lambda_i + \mu_i) P_i &= \lambda_{i-1} P_{i-1} + \mu_{i+1} P_{i+1}, \quad i \geq 1,\end{aligned}\tag{7.17}$$

for which a solution must take the form

$$P_i = \frac{\lambda_{i-1}}{\mu_i} P_{i-1} = \frac{\prod_{k=0}^{i-1} \lambda_k}{\prod_{k=1}^i \mu_k} P_0, \quad i \geq 1.\tag{7.18}$$

The next question is whether or not Equation (7.18) defines a steady state distribution. If so, then we would have

$$1 = \sum_{i=0}^{\infty} P_i = P_0 \left( 1 + \sum_{i=1}^{\infty} \frac{\prod_{k=0}^{i-1} \lambda_k}{\prod_{k=1}^i \mu_k} \right)$$

or equivalently,

$$\begin{aligned}P_i &= \frac{\prod_{k=0}^{i-1} \lambda_k}{\prod_{k=1}^i \mu_k} P_0, \quad i \geq 1, \\ P_0 &= \frac{1}{1 + \sum_{i=1}^{\infty} \frac{\prod_{k=0}^{i-1} \lambda_k}{\prod_{k=1}^i \mu_k}}.\end{aligned}\tag{7.19}$$

Thus, there are two possibilities. If  $P_0 > 0$ , then there exists a steady state distribution. Conversely, if  $P_0 = 0$ , from Equation (7.19)  $P_i = 0$  for any finite  $i$ , and the probability that the process is in state  $i$  approaches 0 in time. This in turn implies that the probability that the process is in any state  $i \leq N$  approaches zero for any finite  $N$ . In this case the process is *idxtransient*, so that the state  $i$  currently occupied by the process tends to increase unboundedly.

We will give an important example of a birth and death process in the next section.

## 7.4 Queueing Systems

Queueing systems form a class of stochastic process of considerable importance in operations research, and present a rich set of applications in control and optimization. They also possess an elegant and intuitive parametric modelling theory, and will therefore serve well to illustrate some of the techniques described in this volume. A queueing system is easy to describe. It contains a queue into which an arrival process of customers enter. It also contains  $m$  servers, who service customers. The time of service has a specified distribution. A customer in the queueing system is either being served, or is in the queue waiting to be served. A server is either busy serving a customer, or is free. Upon service completion the customer exits the system and the server immediately begins service of some customer from the queue if it is not empty. Accordingly, a customer entering the system begins service immediately if there is a free server, or enters the queue otherwise to wait for service. A variety of *queueing disciplines* exist to determine which customer in the queue enters service at the next service completion, the most common being the normally observed FIFO/FCFS discipline (*first in first out* or *first come first served*). The queue discipline affects some system properties (waiting time of an arriving customer) but not others (busy period of server), at least absent additional structure, and so is specified only when relevant.

### 7.4.1 Queueing Systems as Birth and Death Processes

Considerable insight can be gained by modeling queueing systems as birth and death processes. The state variable will be the number of customers in the system (in service and in queue), and either increases by one when a customer arrives, or decreases by one when a service ends. Service times and interarrival times are exponentially distributed (that is, memoryless), but the service time rates  $\mu_i$  and the arrival rates  $\lambda_i$  (using the notation of Section 7.3) are allowed to depend on the state variable  $i$ . The balance equation solution, if it exists, is obtained directly from Equation (7.19). The essential condition is that  $P_0 > 0$ , equivalently, that the sum  $\sum_{i=1}^{\infty} (\prod_{k=0}^{i-1} \lambda_k) / (\prod_{k=1}^i \mu_k)$  is finite. A steady state solution to the balance equations exists if and only if this condition holds.

**Example 7.5 (Single Server Queue)** The simplest queueing model assumes constant arrival and service rates  $\lambda_i \equiv \lambda$  and  $\mu_i = \mu$  with  $m = 1$  server. Define  $\rho = \lambda/\mu$  (this is the utilization factor to be defined in Section 7.4.2). Substitution into (7.19) gives

$$P_0 = \frac{1}{1 + \sum_{i=1}^{\infty} \rho^i} = \begin{cases} 1 - \rho & ; \quad 0 \leq \rho < 1 \\ \infty & ; \quad \rho \geq 1 \end{cases}$$

(Appendix B.5). This means a steady state solution to the balance equations exists if and only if  $\rho < 1$ . If this condition holds, we have from Equation

(7.19)

$$P_i = (1 - \rho)\rho^i, \quad i = 0, 1, \dots$$

Conversely, if  $\rho \geq 1$  the queueing system is transient, and the number of customers in the system will tend to grow unboundedly. It is interesting to note that  $\mu = \lambda$  is not sufficient to ensure the existence of a steady state distribution, that is, the service rate must strictly exceed the arrival rate in order to ensure stability. ■

The birth and death model easily accommodates modifications. Suppose we have  $m > 1$  servers, and each is capable of serving at a rate  $\mu$ . In this case  $\mu_i$  models not a single server rate, but the system service rate. If there are  $i$  customers in service, then by Theorem 7.2 the system service rate is  $i\mu$ . If there are  $i$  customers in the system, there are  $\min(i, m)$  customers in service, so the birth and death process is defined by  $\lambda_i \equiv \lambda$  and  $\mu_i = \min(i, m)\mu$ , and these values may be substituted into (7.19). Similarly, if the queue has finite capacity, that is, it can hold no more than  $K < \infty$  customers, this can be modelled by setting  $\mu_i, \lambda_i$  to zero for all large enough  $i$ . In addition, the number of customers may be a finite number  $M$ . If each potential customer enters the queueing system at rate  $\lambda$ , then the system arrival rate is  $\lambda_i = (M - i)\lambda$  when there are  $i$  customers in the system.

#### 7.4.2 Utilization Factor

The *utilization factor* of a queueing system may be defined as

$$\rho = \lambda/\mu$$

where  $\lambda$  is the average arrival rate of customers, and  $\mu$  is the service rate (see Example 7.5). The precise definition depends on the system, since the arrival and service characteristics need not be time homogenous. In such cases,  $\mu$  may be taken as the maximum service rate. This quantity is fundamental to queueing systems, since we expect the service rate to be higher than the arrival rate, otherwise the queue size will increase indefinitely (this generally holds even when  $\rho = 1$ ). In the single server model this idea is made precise by the observation that the queueing system is stable if and only  $\rho < 1$ . The utilization factor of some importance for many reasons, since even within the constraint  $\rho < 1$  distribution characteristics affecting approximation methods can vary greatly, and computational challenges may arise when  $\rho$  is less than but close to 1.

#### 7.4.3 General Queueing Systems and Embedded Markov Chains

Birth and death queueing system models offer considerable flexibility and insight, but will clearly not be adequate for all systems. As discussed in Section 7.1 the Poisson process will approximate an aggregation of many independent arrival processes, which seems a reasonable assumption for many

actual queueing systems. The assumption that the service time distribution is exponential is more tentative. The exponential density is defined by only one parameter, so that the mean  $\mu$  and standard deviation  $\sigma$  obey a fixed relationship, namely that the *coefficient of variation* is always  $\sigma/\mu = 1$ . There is certainly no reason to think that this value is inherent to service times in any given queueing systems. The coefficient of variation will surely differ significantly between. for example, the time required to process a fixed payment, and the time required for general repair services.

Queueing models are commonly classified using *Kendall's notation* which originally took form  $A/B/m$ , as originally proposed by Kendall in Kendall (1953). It is assumed that the customer arrival process is a renewal process with a renewal distribution described by  $A$ . Then  $B$  refers in the same way to the service time (which are assumed to be independent), and  $m$  is the number of servers. The convention has since been extended, most commonly to  $A/B/m/K/M$ , in which  $K$  is the system capacity, and  $M$  is the number of customers in a finite population. The last two parameters are often omitted when they equal  $\infty$ . The symbols for  $A, B$  are standard, with  $M$  denoting the exponential distribution ( $M$  for 'memoryless'),  $D$  a deterministic, or constant distribution, and  $G$  denoting a general distribution (the assumption of independence is sometimes indicated by the symbol  $GI$ ). Thus, the queueing system of Example 7.5 is an  $M/M/1$  queue.

The queue  $M/M/m/K/M$  may be modeled as a birth and death process, and so the distributional properties can generally be obtained explicitly, with sufficient algebra. If we require a greater variety of distributional properties, this might be done within the context of continuous-time Markov chains by using the 'method of stages', due to A.K. Erlang. For example, we may replace a single exponential service time of rate  $\lambda$  with  $r$  exponential service times in series, each with rates  $\lambda_1, \dots, \lambda_r$ . Thus, completion of service occurs after the sequential completion of  $r$  stages. The resulting distribution of the total service time is referred to as the *Erlangian distribution*,  $E_r$  in Kendall's notation. Note that this system may be modeled as a continuous-time Markov chain, as long as the state space is extended to include the current stage of a customer. There is a considerable variety of density shapes within  $E_r$ , including the gamma distribution. The coefficient of variation can be made arbitrarily small, but can never exceed 1. If  $E_r$  is generated by constructing stages in series, it is also possible to construct stages in parallel. Here, service consists of selecting one of  $r$  stages according to a fixed probability distribution, and then completing service after an exponential waiting time with the rate associated with that stage. This may also be modelled with continuous-time Markov chains, and the resulting service time is referred to as the *hyperexponential distribution*, denoted  $H_r$  in Kendall's notation, and is formally a mixture of exponential densities. These distributions may of course apply also to arrival times. In general, a queue  $G/G/m/K/M$  can be modelled as a continuous-time Markov chain if the arrival time and service times are either  $M$ ,  $E_r$  or  $H_r$ .

Next, consider a  $M/G/1$  queue. In the absence of any Markovian structure a continuous-time Markov chains model cannot be used. A commonly used approach is the *embedded Markov chain* approach Kendall (1953), in which a semi-Markov process is defined by taking transition epochs to be service completions. The state space of the embedded Markov chain  $X_n$  is then interpretable as the number of customers left behind in the system by a departing customer. If  $G$  is the service time distribution, then

$$\begin{aligned} & P\{\text{number of arrivals during service} = k\} \\ &= \alpha_k = \int_0^\infty \frac{(\lambda t)^k}{k!} e^{-\lambda t} dG(t). \end{aligned}$$

The quantities  $\alpha_k$  suffice to determine the transition kernel for  $X_n$ . To see this, suppose  $X_n = 0$ , that is, the system was empty after the  $n$ th service period. At that point the next event must be an arrival, at which point a service period begins immediately. The state  $X_{n+1}$  is determined when this service period ends, and must equal the number of arrivals during this period (since the current customer leaves the system). Thus, the transition distribution is  $P_{0j} = \alpha_j$ , and similar reasoning yields  $P_{ij} = \alpha_{j+i-1}$  for  $i \geq 1$  and  $j \geq i - 1$ . Adjustments may be made for variable arrival rates (for example, with finite capacity  $K < \infty$ ).

Discussion of this equation, and of the  $M/G/1$  more generally, may be found in, for example, Ross (2019) (introductory) or Kleinrock (1975) (advanced).

## Part II

# Statistics - Introduction

## Chapter 8

# Statistical Summaries of Central Tendency

*Statistics* describes the process of making a statement about a *population* of units based on *data* obtained from a *sample* drawn from that population.

In a *random sample* all units of a population have the same probability of being selected, and the probability that a specific unit is included in the sample does not depend on any other selection. Alternatively, a sample is random if all possible samples of size  $n$  have equal probability of being selected.

If we *sample with replacement*, a unit may be selected more than once. In effect, a selected unit is returned to the population, and may be selected again.

If we *sample without replacement*, a unit may not be selected twice. In effect, a selected unit is removed from the population.

The process of making a statement about a population based on a sample is generally referred to as *inference*. A complete statistical statement must say something about the accuracy of a statement, in which case we may refer to *statistical inference*.

Most statistical statements are one of two kinds.

1. An *estimate* is a quantity calculated from the sample which is expected to be close to some numerical attribute of a population.
2. A *hypothesis* is a statement about a population. The sample is used to decide whether or not the hypothesis is correct.

**Example 8.1** A poll states that 37% of the population believe that water should not be fluoridated. The margin of error is  $+/- 3\%$  19 times out of 20. This means that for surveys using this methodology the probability that the reported proportion is within the margin of error of the true proportion is 19/20. ■

As an example of an *estimate*, suppose randomly chosen sample of families are asked how many children they have. The answers are averaged. It

is assumed that the resulting average is close to the true average number of children per family.

As an example of a *hypothesis*, suppose this family survey was done on 1990 and again in 1999. It is believed that the average family size has decreased. When the two sample averages are compared it is found that the average family size from the 1999 sample is in fact smaller than that of the 1990 sample. This is taken as evidence that family size has decreased.

We next consider the distinction between sampling *with* and *without* replacement.

**Example 8.2** If we did a telephone sample, and we happened to phone the same person twice, would this respondent be represented twice in the sample? If the size of a finite population was 500, the probability that this happens for a sample of size 10 is approximately 8.7% (we'll see how this can be calculated in subsequent sections). ■

The distinction between sampling with and without replacement can lead to some inconsistency. In many practical cases it is natural to think of sampling as being *without replacement*. If this is because the population is large enough that the probability of sampling the same unit twice is very small (but not zero) then there is little practical difference between sampling with and without replacement. However, if we are using a sampling protocol that prohibits repeated sampling of any unit, then the difference may be important. In this case it is important to note that most probability theory used to support inference statements is based on the assumption that sampling is *with replacement* (for more on this subject see, for example, Chapter 22 of Pagano and Gauvreau, 2000).

**Example 8.3** To continue Example 8.2, if the size of a finite population is now 5000, the probability of repeated sampling of any unit in a sample of size 10 is approximately 0.9%. For a finite population size of 50,000, this probability becomes 0.09%. ■

## 8.1 The Role of Variation in Statistics

**Example 8.4** A family has 5 children. Four are male. Is this unusual?

For convenience, we will assume throughout this example that we are dealing with families of exactly 5 children. To answer the question, it might be useful to list all possible male/female combinations:

```
FFFFF
FFFFM FFFMF FFMFF FMFFF MFFFF
```

FFFMM FFMFM FMFFM FFMMF FMFMF MFFMF FMMFF MFMFF MMFFF  
 MMMFF MMFMF MFMMF FMMMF MMFFM MFMFM FMMFM MFFMM FMFMM FFMM  
 MMMMF MMMFM MMFMM MFMMM FMMMM  
 MMMMM

In this case there are 32 possible combinations. In the diagram, they are separated into rows according to how many female children there are. Note that there are 5 configurations with exactly 1 female and 6 configurations with no more than 1 female. Therefore, there is a  $6/32 = 18.75\%$  chance that there will be no more than one female child.

Question: Is the relevant probability  $5/32$  (= the proportion of families with *exactly* 1 female child), or  $6/32$  (= the proportion of families with *no more than* 1 female child)?

We can also say that there is a  $12/32 = 37.5\%$  chance that there will be either at most one female or at most one male child (or, that one of the genders will be represented by at most one child). We can say that this particular configuration is not unusual.

Question: Is the relevant probability  $6/32$  (= the proportion of families with *no more than* 1 female child) or  $12/32$  (= the proportion of families with *no more than* 1 child of one or the other gender)?

We will revisit this type of question in subsequent sections. ■

**Example 8.5** As a somewhat artificial example suppose that in some small village the number of families is 5 [!], and that we select at random a sample of 4. Assume that the true numbers of children are

$$1, 3, 4, 5, 8$$

There are 5 ways to select a sample of size 4 (since there are 5 ways to leave one family out of the sample). The five possible means are

$$\begin{aligned} \frac{1+3+4+5}{4} &= 3.25 \\ \frac{1+3+4+8}{4} &= 4 \\ \frac{1+3+5+8}{4} &= 4.25 \\ \frac{1+4+5+8}{4} &= 4.5 \\ \frac{3+4+5+8}{4} &= 5 \end{aligned}$$

Note that the resulting averages range from 3.25 to 5, depending on the particular sample chosen. A crucial step in statistical inference is the measurement of the variability inherent in calculating averages, or other quantities, from samples. ■

The first step in most statistical analyses is to construct a summary of the data in a sample. There are two kinds of summaries, namely *numerical summaries* and *graphical summaries*. The main goal of a summary is to display some essential feature of the data.

We'll introduce some notation that will be important throughout these notes. We will sometimes denote a data set (or, more informally a sample) by the notation

$$X_1, X_2, \dots, X_n.$$

This represents a general indexed list of  $n$  numbers. If we have a list of four numbers 10, 5, 7, 5 then in the above notation  $X_1 = 10$ ,  $X_2 = 5$ ,  $X_3 = 7$  and  $X_4 = 5$ .

We first consider the problem of summarizing the *central tendency* of the data, in effect, determining a number which is in some sense ‘typical’ of the values in the data set.

## 8.2 The Sample Mean

A sample mean is an average of a sample of some measurements. The general formula is

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}.$$

The notation

$$\sum_{i=1}^n$$

is shorthand for “sum the following indexed list for  $i = 1$  to  $i = n$ ”. If we only wanted to sum the first three items in the list we could write

$$\sum_{i=1}^3 X_i.$$

**Example 8.6** If  $X_1 = 10$ ,  $X_2 = 5$ ,  $X_3 = 7$  and  $X_4 = 5$ , then  $n = 4$ , so that

$$\begin{aligned}\bar{X} &= \frac{\sum_{i=1}^n X_i}{n} \\ &= \frac{\sum_{i=1}^4 X_i}{4} \\ &= \frac{X_1 + X_2 + X_3 + X_4}{4} \\ &= \frac{10 + 5 + 7 + 5}{4} \\ &= 6.75\end{aligned}$$

■

Using R we have the `mean()` function:

```
> x = c(10, 5, 7, 5)
> mean(x)
[1] 6.75
>
```

If there is missing data we can use the `na.rm = TRUE` option. The missing data is, of course, not included in the determination of  $n$ :

```
> x = c(10, 5, 7, 5, NA)
> mean(x, na.rm=T)
[1] 6.75
```

The sample mean is perhaps the most widely used statistic. However, it is often used inappropriately, as the following examples show.

**Example 8.7** Suppose on a major league baseball team, the pitching staff has the following yearly salaries, in \$1,000's of dollars

150	150	270	340	375
560	605	1,030	1,100	6,500

The average salary here is \$1,108,000. This is larger than all but the highest salary of \$6,500,000. Although the word *average* is often used interchangeably with *typical* the average here is far from typical. The problem is that the average is very sensitive to extremely large (or small) values. Here, the largest value of 6,500 is some much larger than the other values that has a disproportionate effect on the mean. Later on, we will discuss some ways to remedy this situation. ■

### 8.3 Order Statistics

Given a data set  $X_1, X_2, \dots, X_n$ , there will often be the need to consider the *order* of the observations. Usually, the notation used is:

$$X_{(k)} = k\text{th highest observation among } X_1, X_2, \dots, X_n.$$

The *order statistics* are then

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}.$$

**Example 8.8** Suppose

$$X_1 = 2.4, X_2 = 9.5, X_3 = 2.4, X_4 = -10.4.$$

Then

$$X_{(1)} = -10.4, X_{(2)} = 2.4, X_{(3)} = 2.4, X_{(4)} = 9.5. \quad \blacksquare$$

Note that ties do not alter the definition of the order statistics. ■

## 8.4 The Median

The median is the “middle” value of a data set. To calculate the median, first sort the data.

The sorted values of

$$5, 6, 2, 1, 4, 2, 3$$

are

$$1, 2, 2, 3, 4, 5, 6.$$

There are seven numbers, so the middle value is the 4th highest. That is, the median is 3.

Suppose there are an even number in the list. We then take the average of the middle two numbers.

If we add a seven to the list we have

$$1, 2, 2, 3, 4, 5, 6, 7$$

in which case the middle two numbers are the 4th and 5th highest, 3 and 4. The median is then the average of 3 and 4, namely 3.5.

**Definition 8.1** Suppose

$$X_1, X_2, \dots, X_n$$

is a list of numbers. Suppose

$$X_{(1)}, X_{(2)}, \dots, X_{(n)}$$

are the order statistics (Section 8.3). Let  $\tilde{X}$  be the median:

1. If  $n$  is odd, then  $\tilde{X} = X_{(\frac{n+1}{2})}$ ,
2. If  $n$  is even, then  $\tilde{X} = \frac{X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}}{2}$ .

■

In R, use the command `median()`:

```
> x = c(11, 30, 32, 51, 57, 60)
> median(x)
[1] 41.5
> (32+51)/2
[1] 41.5
>
```

One advantage of the median over the mean is that it is less sensitive to extreme values which would have a large effect on the mean.

**Example 8.9** To return to the baseball example, note that there are  $n = 10$  values, so that the median is given by

$$\begin{aligned}\tilde{X} &= \frac{X_{(5)} + X_{(6)}}{2} \\ &= \frac{375 + 560}{2} \\ &= 467.5\end{aligned}$$

or, the median salary is \$467,500, which is considerably more typical than \$1,108,000. To emphasize the point that the median is less sensitive to extreme values than the mean, consider the effect of removing the salary valued \$6,500,000 from the sample.

	Complete sample	Largest value removed	Percent change
Mean $\bar{X}$	\$1,108,000	\$508,889	-54.1%
Median $\tilde{X}$	\$467,500	\$375,000	-19.8%

Removing the largest values results in a reduction of the sample mean of more than 50%, and in a reduction of the median of only 20%. ■

## 8.5 The Trimmed Mean

The  $K\%$  trimmed mean of a sample  $X_1, X_2, \dots, X_n$  is calculated by first removing the highest  $K\%$  and the lowest  $K\%$  of the data (or as closely as possible to  $K\%$ ), then calculating the sample mean of the remaining data.

**Example 8.10** If we have a sample

$$10, 12, 22, 12, 23, 13, 15, 30, 14, 3$$

we first sort it to get

$$3, 10, 12, 12, 13, 14, 15, 22, 23, 30.$$

To calculate the 10% trimmed mean, note that  $n=10$  so that the highest and lowest 10% of the data represent simply the highest and lowest single value. We then remove the 3 and the 30. The average of the remaining values is then

$$\begin{aligned}\bar{X}_{10\%} &= \frac{10 + 12 + 12 + 13 + 14 + 15 + 22 + 23}{8} \\ &= 15.125.\end{aligned}$$

■

Note that both the sample mean and the median are special cases of the trimmed mean, derived by setting, respectively,  $K = 0\%$  and  $K = 50\%$ .

**Example 8.11** The trimmed mean can be calculated in R using the `trim` option of the `mean` function. Making use of the data from Example 8.7 we have:

```
x = c(150, 150, 270, 340, 375, 560, 605, 1030, 1100, 6500)
> mean(x, trim = 0.1)
[1] 553.75
```

We'll update the previous baseball salary table of Example 8.9 by including the 10% trimmed mean:

	Complete sample	Largest value removed	Percent change
Mean $\bar{X}$	\$1,108,000	\$508,889	-54.1%
10% Trimmed Mean $\bar{X}_{10\%}$	\$553,750	\$475,714	-14.1%
Median $\tilde{X}$	\$467,500	\$375,000	-19.8%

The trimmed mean is a type of compromise between the mean and the median, accordingly, its value is between the mean and median. Note also that the trimmed mean has changed the least after moving the highest salary. This is a reflection of the fact that the trimmed mean tends to be more stable, or less variable, than the median. ■

## 8.6 Sample Quantiles and Percentiles

Recall the definition of a quantile of a RV with CDF  $F_X$  (Definition 4.4 from Section 4.4). Just as we have sample analogs  $\bar{x}$  and  $S^2$  which estimate population parameters  $\mu$  and  $\sigma^2$ , we may construct a *sample quantile*  $\hat{Q}(p)$  from data to estimate a  $p$ -quantile.

The sample median is a special case of a sample quantile,  $\hat{Q}(0.5)$  in particular. Intuitively, the sample median is a number below which  $1/2$  of the data are located and above which  $1/2$  of the data are located (or as nearly as possible). The *sample p-quantile* is the value below which a proportion  $p$  of the data are located and above which  $1 - p$  of the data are located. Similarly, the sample  $K$ th percentile is the sample  $(K/100)$ -quantile.

Sample quantiles can be constructed from the order statistics (Section 8.3), as shown in the following example.

**Example 8.12** Suppose out of a class of 150, a scholarship is to be given to those students with a mark above the 95th percentile. If the marks are given below above which mark are scholarships awarded?

```
40.1 40.7 41.7 42.2 42.4 43.1 43.1 43.5 43.5 44.2
44.3 44.4 44.8 45.1 45.3 45.4 45.5 46.4 46.4 47.0
```

```

47.1 47.4 47.5 48.1 48.9 49.3 49.3 49.4 49.4 49.5
49.8 49.8 49.9 50.0 50.5 51.6 52.7 52.8 53.2 53.6
54.8 55.6 56.5 56.6 57.4 57.8 57.8 58.0 58.4 58.7
59.0 59.0 59.1 59.8 59.9 60.0 60.4 60.4 60.6 60.7
60.8 61.1 61.7 61.8 62.2 62.3 62.3 62.4 63.2 64.0
64.5 64.6 64.7 65.3 65.4 66.3 66.5 67.2 67.4 67.5
68.2 68.8 69.1 69.4 69.6 69.7 69.8 69.9 69.9 70.5
70.5 71.0 71.2 71.5 71.8 71.9 72.1 72.8 72.8 73.5
73.6 73.9 73.9 73.9 74.9 76.1 76.3 76.6 76.7 77.2
77.9 78.2 78.8 78.9 79.0 80.6 80.7 80.8 83.1 83.4
83.9 84.2 84.4 84.6 85.4 85.4 85.6 85.7 85.8 86.4
86.5 86.9 89.4 89.6 89.9 90.0 90.0 90.5 90.9 91.0
91.2 91.5 91.7 92.2 92.2 92.7 92.8 93.6 93.9 94.3

```

Now, 95% of 150 is  $0.95 \times 150 = 142.5$ . If we select a mark in between the 142nd and 143rd highest (that is, in between  $X_{(142)}$  and  $X_{(143)}$ ) we will have an approximate 95th percentile. From the table, these are 91.5 and 91.7. We therefore set the 95th percentile to be 91.6, although this number does not uniquely satisfy the definition of the 95th percentile. ■

In R sample quantiles are calculated using the `quantile()` function. Usually, we cannot find a single number that satisfies the definition of a quantile. For example, we cannot precisely define a 10% percentile of 7 numbers.

Suppose we are given a continuous CDF  $F_X(x)$ . Recall from Section 4.4 the quantile function  $Q(p) = F_X^{-1}(p)$ . Sample quantiles can be approximated by estimating a quantile function from the sample. R provides 9 methods for doing this, specified by setting the `type` option to be an integer from 1 to 9. The choice of method depends on whether the sample is assumed to come from a discrete distribution (types 1-3) or a continuous distribution (types 4-9). The default is `type = 7`. For this method the quantile function for  $n$  data points is estimated by linear interpolation between points:

$$(p[k], x[k]), \quad k = 1, \dots, n,$$

where  $p[k] = (k - 1)/(n - 1)$  and  $x[k]$  is the  $k$ th largest data value.

Method type 6 is similar to type 7, except that  $p[k] = k/(n + 1)$ . The choice of method can make a big difference with smaller sample sizes, should give similar values for larger sample sizes:

```

> x = c(150, 150, 270, 340, 375, 560, 605, 1030, 1100, 6500)
> quantile(x, probs = c(0.25, 0.5, 0.75))
  25%    50%    75%
287.50 467.50 923.75
> quantile(x, probs = c(0.25, 0.5, 0.75), type = 7)

```

```

25%      50%      75%
287.50  467.50  923.75
> quantile(x, probs = c(0.25, 0.5, 0.75), type = 6)
25%      50%      75%
240.0   467.5   1047.5
> median(x)
[1] 467.5

```

Note that multiple quantiles can be requested in a single function call. The above example requests the 25%, 50% and 75% percentiles.

Hyndman and Fan (1996) can be recommended as a survey on this topic.

## 8.7 Random Samples and Data Homogeneity

The most common type of statistical analysis concerns the *random sample*. One important issue in statistical analysis, often overlooked, is that of *homogeneity*. An implicit assumption in many analyses is that the units of a population are of some single type that is the object of study (that is, the population is *homogeneous*).

When can we say that there is more than ‘one type’ of unit in a population? When we can identify multiple types, the relative frequencies of which significantly affect the quantitative properties of the population (in which case the population is *heterogeneous*).

**Example 8.13** Suppose a company has the following paid positions

Position	Number	Salary	Total Payroll
Laborer	3	25,000	75,000
President	1	60,000	60,000
Total	4		135,000
Average Salary			33,750

The average salary of this company (of 4 salaries) is \$33,750. Now suppose that things improve for this company. Two new laborers are hired, and all positions receive a \$2,500 raise. The table now looks like this:

Position	Number	Salary	Total Payroll
Laborer	5	27,500	137,500
President	1	62,500	62,500
Total	6		200,000
Average Salary			33,333

Curiously, the average salary has gone down, even though all salaries have increased. This is because the proportion of lower paid positions has increased, offsetting the raises. The fundamental mistake here is that the measurements making up the average are nonhomogeneous (also known as “comparing apples and oranges”). ■

## Chapter 9

# Graphical Summaries

We have seen a number of numerical summaries of data in the previous section. The sample mean, median, and trimmed mean were measures of *central tendency* (we'll revisit this term below). Sometimes, this type of summary is used to characterize a ‘typical’ number from a set of data (ages, incomes, sizes, and so on).

However, we already know that we must be careful in using a summary of central tendency to describe what is ‘typical’. For example, from the baseball salary example we know that ‘average’ and ‘typical’ need not mean the same thing.

We need to consider more precisely why, or why not, a summary of central tendency may be a good representation of what is ‘typical’. We could say that a summary of central tendency represents a typical value of a data set if about 1/2 the data are above the summary, and 1/2 the data are below. By its construction, the median satisfies this definition of ‘typical’.

However, following the baseball salary example, if we wished to extrapolate the summary, that is, use the summary to estimate the total payroll of a baseball team, we would prefer to use the sample mean.

We could also say that a summary of central tendency is a good representation of what is ‘typical’ if most data values are near it. Unfortunately, we have no way of knowing from the value any single summary of central tendency whether or not it is ‘typical’ in this sense.

In order to resolve these problems (and for many other reasons) we need to view the data as a whole, and this is done using *graphical summaries*.

### 9.1 Stem and Leaf Plots

As an example of a graphical summary, we now introduce the stem and leaf plot. Consider the following sorted data set

1023	1044	1151	1186	1206	1215	1241	1244	1270	1292
1302	1324	1337	1374	1435	1478	1549	1590	1703	1762

One way to do this is to take each number and divide it into two parts, the *stem* and the *leaf*.

$$\begin{aligned} \text{stem} &\Rightarrow 10|23 \Leftarrow \text{leaf} \\ \text{stem} &\Rightarrow 10|44 \Leftarrow \text{leaf} \\ &\vdots \\ \text{stem} &\Rightarrow 17|62 \Leftarrow \text{leaf} \end{aligned}$$

The choice of stem and leaf is a matter of judgement, but the dividing line must be between the same two digits for each number.

The stem and leaf plot is then constructed by making a column consisting of all stems present. Then for each number, take the first digit of the leaf and place it in the row labeled by the appropriate stem.

10 24
11 58
12 014479
13 0237
14 37
15 49
16
17 06

The leafs in the plot are usually sorted. The number 1324 is represented by the leaf 2 in the row labeled with stem 13. Note that there are no gaps in the stems. Even though there are no numbers in the data set starting with 16, we include the stem 16 in the plot without placing any leafs next to it.

The point of the plot is to give a quick impression of the data. We can tell, for example, that the numbers tend to be centered within the 1200 to 1400 range, then tail off above and below.

Next, consider the following data set

832	833	835	836	836	837	837	838	839
840	841	841	842	843	843	844	845	847
850	853	854	855	855	856	856	857	858
858	859	860	861	861	862	863	863	868

It seems pointless to choose the first digit as the stem, since in that case we'd have only one stem valued 8. If we use the first two digits as the stem we get the following stem and leaf plot:

83 235667789
84 011233457
85 03455667889
86 0112338

There seems to be too few stems to give an informative plot. One alternative is to breakdown each row into two rows, with leafs numbered 0-4 in one and leafs numbered 5-9 in the next. This gives

83 23
83 5667789
84 0112334
84 57
85 034
85 55667889
86 011233
86 8

Now, the plot has more shape. We can see, for example, that there are two peaks, or points at which values tend to be concentrated.

## 9.2 Histograms

The histogram is a device which is similar to the stem and leaf plot, but which has more flexibility. Suppose we have a data set with 100 observations, ranging from 11.1 to 29.6. Figure 9.1 is a *histogram* of the data.

First notice that we can associate numbers from 10 to 30 with the bars. Next note that there is a vertical axis. The histogram tells us that there are 6 of the 100 observations near the number 15. In particular there are 6 observations between the numbers 14.5 and 15.5. Also, there are 2 observations between 10.5 and 11.5, but no observations between 11.5 and 12.5.

We'll now go through the steps of constructing a histogram:

1. Determine the minimum and maximum of the data set.
2. Define a series of *class intervals*. These are numerical intervals which are
  - (a) Equal in size, and
  - (b) Adjacent to each other.

The first class interval should contain the minimum and the last class interval should contain the maximum. If an observation is exactly on the boundary between two class intervals, it goes in the higher one.

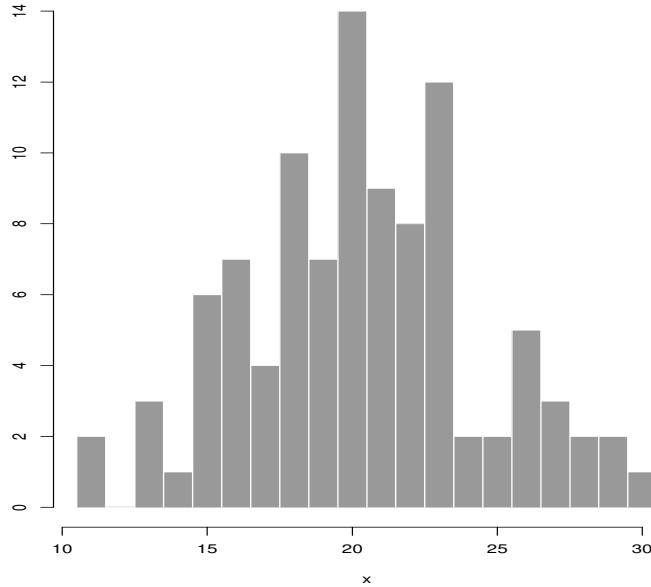


Figure 9.1: Example of a histogram.

3. For each class interval calculate the exact number of observations which fall into it. This is the class frequency.
4. Indicate the position of the class intervals on a horizontal axis. This is usually done by indicating the position of the midpoint of the interval, as in Figure 1.
5. At each class interval draw a vertical bar equal in height to the class frequency.

There are two other ways to set the vertical axis. One alternative is to calculate the *relative class frequencies*. These are the *proportions* of observations in each class interval. Again, the vertical bar is drawn proportional in height to the relative class frequency.

The other alternative is to set the bars equal in height to the *density*, which is defined as the relative frequency divided by the class interval width. In fact, if for some reason it is advantageous to have class intervals of different widths, then using the density is the preferable method. A useful mathematical fact about the density is that the total area taken up by the bars of the histogram always sums to 1.

**Example 9.1** We'll construct a histogram of the following 20 observations.

100.5	101.4	103.5	105.3	105.9	106.2	107.0	108.1	108.7	108.8
108.9	109.5	111.3	114.1	115.9	116.4	116.7	116.7	117.3	118.1

The minimum value is 100.5 and the maximum is 118.1. We'll define the following 5 class intervals.

Class	Start	End	Class	Relative Class	Density
			Frequency	Frequency	
1	100	103.9	3	0.15	0.0375
2	104	107.9	4	0.20	0.05
3	108	111.9	6	0.30	0.075
4	112	115.9	2	0.10	0.025
5	116	120.0	5	0.25	0.0625

The intervals were obtained by dividing the interval 100 to 120 into 5 equal spaces. After we draw the bars we have the histogram in Figure 9.2.

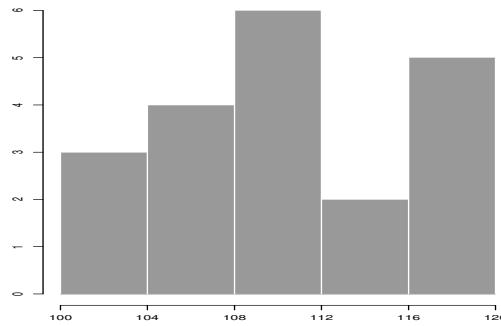


Figure 9.2: Example of histogram.

If we were to use the relative class frequencies we would have the histogram of Figure 9.3.

Note that the two histograms have exactly the same shape. Only the vertical axis is different. ■

### 9.2.1 Creating Histograms in R

The main R function for creating histograms is `hist()`. It can often be used in the form `hist(x)`, where `x` is usually a vector, but can also be a matrix. There are a number of options. Perhaps the most important of these of the `breaks` option and the `nclass` option. Without these `hist()` will use a default algorithm to determine the number of classes and the breakpoints, but these can be supplied by the user. Consult `help(hist)` for details:

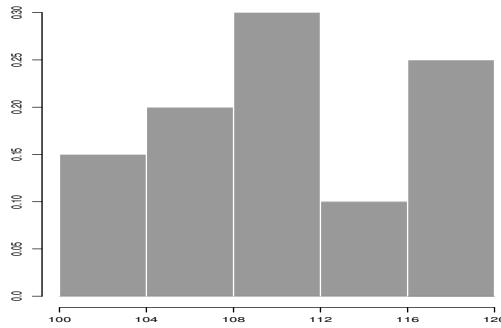


Figure 9.3: Histogram for Example 9.1.

```

> par(mfrow=c(2,3))
> x = rnorm(100, mean = 200, sd = 0.1)
> hist(x, main = "Normal Distribution, mean = 200, SD = 0.1, n = 100")
> x = rnorm(10000, mean = 200, sd = 0.1)
> hist(x, main = "Normal Distribution, mean = 200, SD = 0.1, n = 10000")
> hist(x, nclass=100,
      main = "Normal Distribution, mean = 200, SD = 0.1, n = 10000")
> x = runif(100)
> hist(x, main = "Uniform Distribution")
> x = rexp(100, rate = 5)
> hist(x, main = "Exponential Distribution, rate = 5")
> x = rpois(100, lambda = 100)
> hist(x, main = "Poisson Distribution, lambda = 100")
>

```

Note the function `par()`. This is a very powerful function, used to define the properties of the graphical device. In this case, the option `mfrow` is a vector setting the number of rows and columns of plots which may be displayed.

### 9.3 Boxplots

The *boxplot* is a useful (if not immediately intuitive) graphical device. To construct a boxplot we need the following quantities.

1. The median ( $Q_2$ )
2. The *lower quartile* ( $Q_1$ ) and *upper quartile* ( $Q_3$ ), which are the 25th and 75th percentiles. (The median is then called the *middle quartile*).

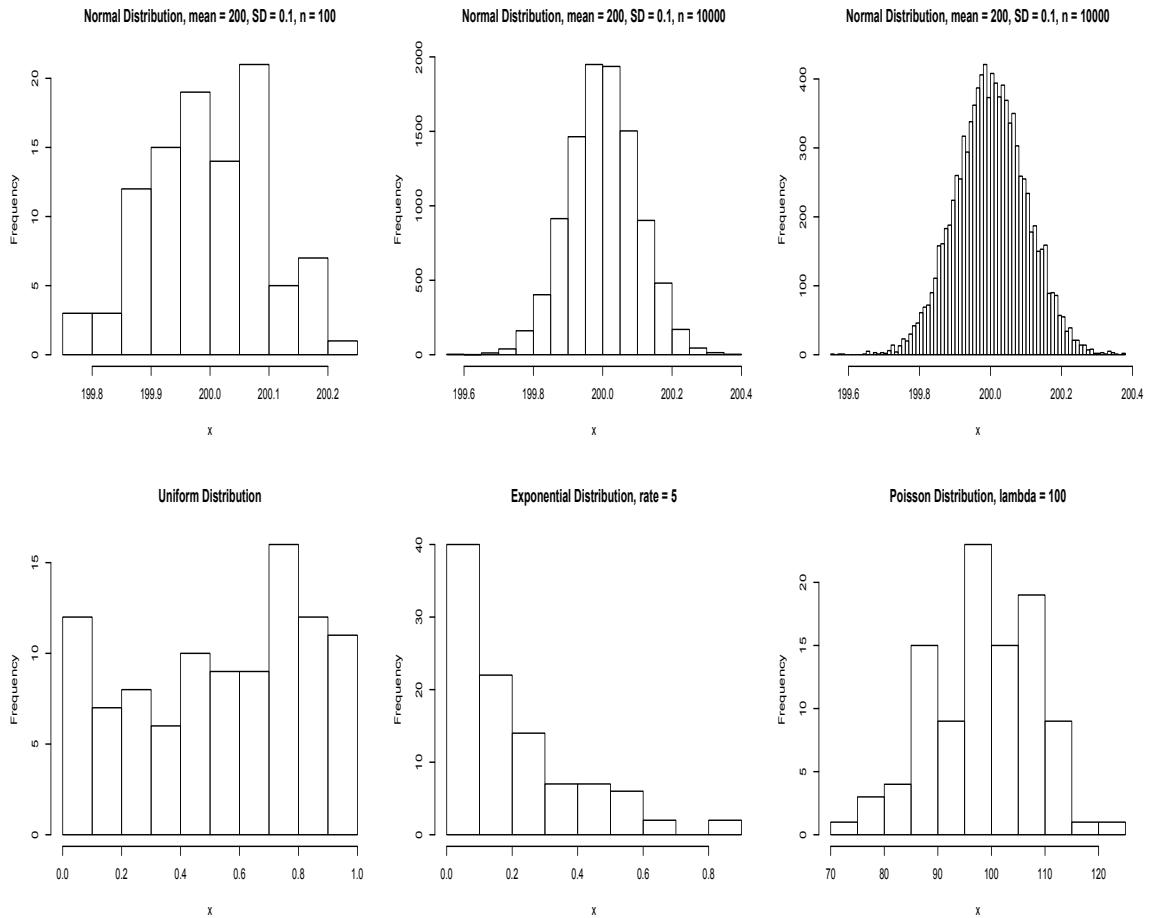


Figure 9.4: Creating histograms in R

3. The *interquartile range* is given by the formula

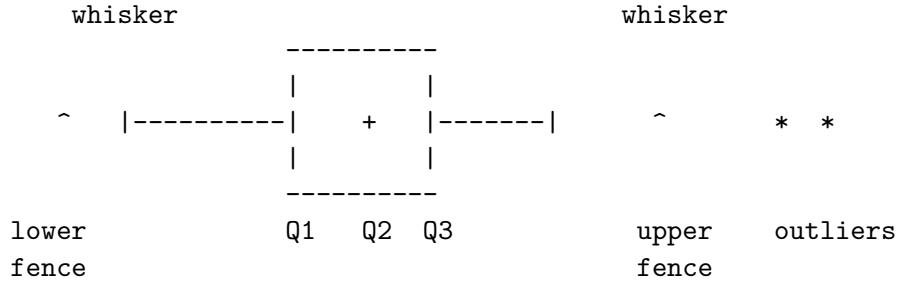
$$\text{IQR} = Q_3 - Q_1$$

4. The standard span is  $1.5 \times \text{IQR}$ .

To construct the boxplot do the following steps.

1. Construct an axis covering all observations.
2. Draw a box with ends at  $Q_1$  and  $Q_3$ .
3. Indicate the location of the median inside the box with a line or other symbol.

4. Locate the *upper fence*, obtained by adding the standard span to  $Q_3$ .
5. Locate and indicate on the plot the *whisker*, given by the largest observation that is lower than the upper fence. Draw a line from the box to the whisker.
6. Locate the *lower fence*, obtained by subtracting the standard span from  $Q_1$ .
7. Locate and indicate on the plot the other whisker, given by the smallest observation that is larger than the lower fence. Draw a line from the box to the whisker.
8. Indicate individually any observations which are beyond the fences. Such observations are usually called *outliers*



**Example 9.2** Suppose we have observations

8.5	14.7	34.7	37.6	40.6	42.5
43.6	48.0	51.9	54.5	57.7	60.5

There are  $n = 12$  numbers, so the median is the average of the 6th and 7th highest observation, which is

$$\begin{aligned} Q_2 &= \frac{X_{(6)} + X_{(7)}}{2} \\ &= \frac{42.5 + 43.6}{2} \\ &= 43.05 \end{aligned}$$

Note that 25% of 12 is 3, so the 25th percentile can be taken as the number in between the 3rd and 4th largest observation,

$$\begin{aligned} Q_1 &= \frac{X_{(3)} + X_{(4)}}{2} \\ &= \frac{34.7 + 37.6}{2} \\ &= 36.15 \end{aligned}$$

Also, 75% of 12 is 9, so the 75th percentile can be taken as

$$\begin{aligned} Q_3 &= \frac{X_{(9)} + X_{(10)}}{2} \\ &= \frac{51.9 + 54.5}{2} \\ &= 53.2. \end{aligned}$$

This gives

$$\begin{aligned} \text{IQR} &= Q_3 - Q_1 \\ &= 53.2 - 36.15 \\ &= 17.05 \\ \text{standard span} &= 1.5 \times \text{IQR} \\ &= 1.5 \times 17.05 \\ &= 25.575 \\ \text{upper fence} &= Q_3 + \text{standard span} \\ &= 53.2 + 25.575 \\ &= 87.775 \\ \text{lower fence} &= Q_1 - \text{standard span} \\ &= 36.15 - 25.575 \\ &= 10.575 \end{aligned}$$

To complete the calculations, we note that there are no observations above the upper fence, and the largest value less than the upper fence is 60.5, which gives the upper whisker. On the other hand, note that there is one value below the lower fence, namely 8.5, which is therefore an outlier. The smallest value above the lower fence is 14.7, which is the lower whisker. This results in the following boxplot (Figure 9.5). ■

**Example 9.3** Boxplots are particularly useful when more than one are placed on the same axis for purposes of comparison. A local study measured the survival times from diagnosis of prostate cancer of 84 men. The subjects were divided into two groups. Group one consists of those men who were more than 65 years old at diagnosis, and group two consists of those aged 65 or less at diagnosis. A boxplot of survival time in years was constructed for each group and placed on the same axis (Figure 9.6). As can be seen, there is a tendency for those in group two to have longer survival times. However, there is considerable variation within the two groups, so that age by itself may be a poor predictor of prognosis. ■

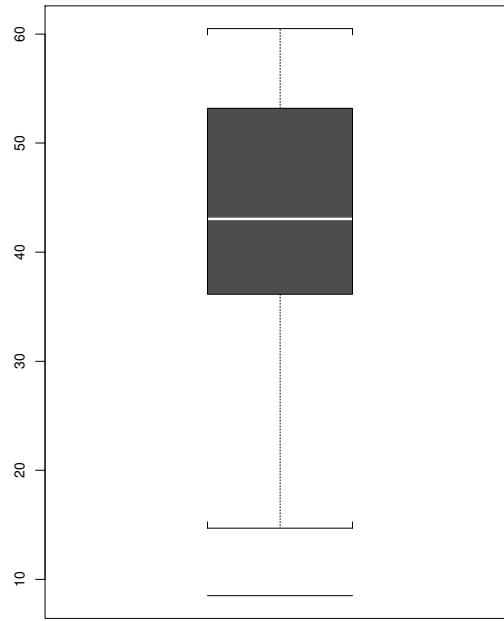


Figure 9.5: Boxplot of Example 9.2.

### 9.3.1 Creating Boxplots in R

The main command in R for boxplots is `boxplot()`. As for `hist()`, boxplots can be created easily by inputting a single vector, that is `boxplot(x)`. However, a boxplot is a somewhat more complicated device than the histogram, in the sense that it is often used to compare data sets. Multiple data sets can be introduced within a list, or, when appropriate, as a matrix, in which case the data is separated by column:

```
> par(mfrow=c(1,2))
> x = rnorm(100, mean = 10, sd = 2)
> y = rnorm(200, mean = 15, sd = 1)
> boxplot(list(x,y), names = c("Sample 1", "Sample 2"),
           main="Two Boxplots")
>
> m = matrix(rnorm(100),10,10)
> boxplot(m, main = "Column-wise boxplots for matrix input")
```

The resulting boxplots are shown in Figure 9.7.

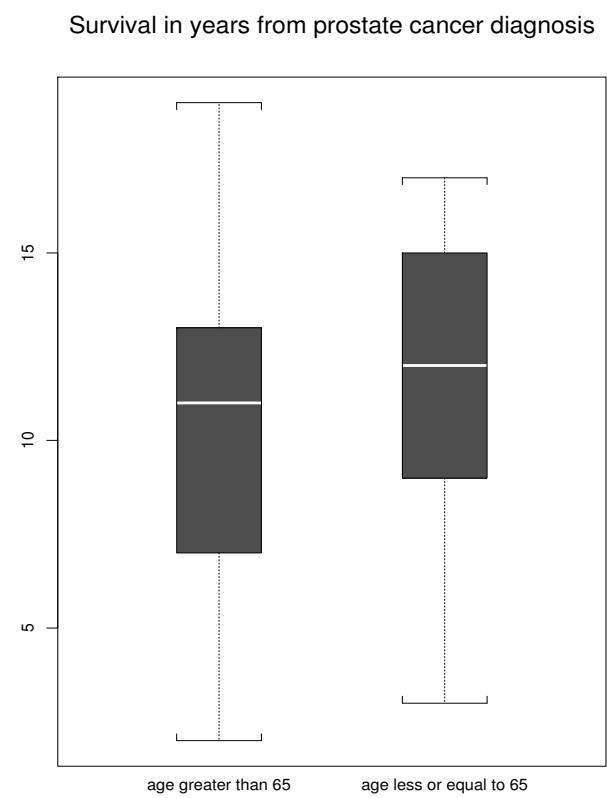


Figure 9.6: Boxplots of survival times for prostate cancer study for Example 9.3.

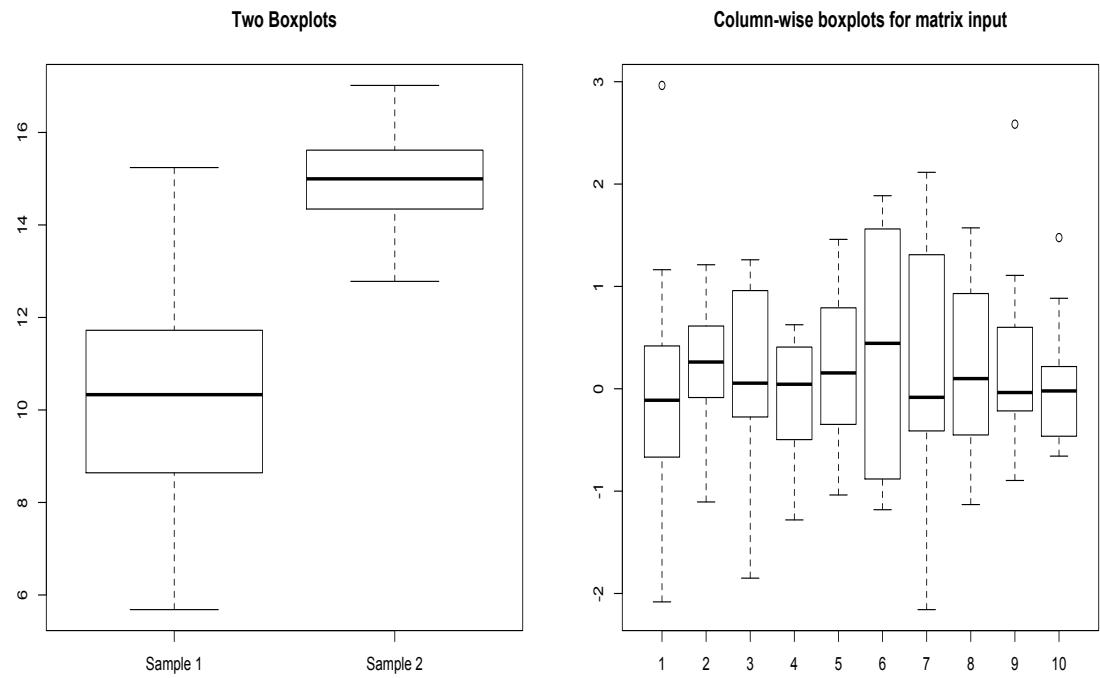


Figure 9.7: Creating boxplots in R , Section 9.3.1.

# Chapter 10

# Properties of Data

Collections of measurements have properties which arise from the relative values of the individual measurements. In this section we explore more precisely which of these properties are important, and how to quantify and analyze them.

## 10.1 Types of Data

Data can be classified into two broad categories:

1. *Categorical Data.* Data is categorical if each measurement assumes one of several categorical values, which have no numerical significance. A common example is blood type, in which each measurement is one of A, B, AB or O. If there are only two categories the measurements are often called *dichotomous*. Categorical data can be further broken down into two types:
  - (a) *Nominal Data.* If there are no relationships between the categories then the measurements are called *nominal*. Blood types are one example.
  - (b) *Ordinal Data.* If the categories can be meaningfully ranked, then the measurements are called *ordinal*. For example, if the categories are "nonsmoker", "light smoker" and "heavy smoker", it makes sense to assign ranks to the measurements, even though the measurements are not strictly numeric.
2. *Numerical Data* Data are numerical if they are represented by numbers that can be meaningfully ranked and on which it is meaningful to perform arithmetic operations.

- (a) A *discrete* measurement is one that assumes one value from a list of values. The list may be finite or infinite. The most common type of discrete measurement is a count which assumes values  $0, 1, 2, \dots$ .
- (b) A *continuous* measurement is one that can assume any value between two specific values (or all possible numbers). This is commonly used for measurable quantities such as weight, height, speed, and so on.

Note that categorical data are often coded using numbers for convenience, but this alone does not make them numerical observations.

## 10.2 Distributions

A *distribution* is a property of a collection of measurements which describes the frequency at which various measurements occur. Suppose we have 10 measurements

1    1    1    1    2    2    2    3    3    4.

A complete description of the distribution of the measurements can be given by noting that 1 occurs 40% of the time, 2 30%, 3 20% and 4 10%.

The situation is a little more complex if the data are continuous. If we have a data set

1.1    1.3    1.4    1.9    2.3    2.4    2.8    3.4    3.5    4.6.

there is little purpose in noting that 10% of the data are 1.1, 10% of the data are 1.3, and so on. It would be more useful to note that 40% of the data are in the range [1.0,2.0], 30% of the data are in the range [2.0,3.0], 20% of the data are in the range [3.0,4.0] and 10% of the data are in the range [4.0,5.0]. Note that in order to make this comparison meaningful, the ranges should be of equal length. This description of the density tells us that an observation is more likely to be at the lower end of the observed range (near 1.0) than at the higher end (over 4.0).

A distribution is well represented by a histogram. The above data has the following histogram shown in Figure 10.1.

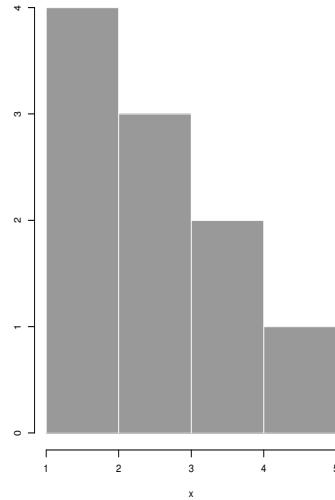


Figure 10.1: Example of histogram for Section 10.2.

As a theoretical device, a collection of measurements can be represented by a *density function*. The density function is a function with the following properties.

1. The function is always greater than or equal to zero.
2. The total area under the function is one.
3. The total area under the function between two numbers  $a$  and  $b$  is equal to the proportion of measurements between  $a$  and  $b$ .

Figure 10.2 has a density function superimposed on a histogram.

### 10.3 Central Tendency and Variability

Usually, the two most important properties of a distribution are the central tendency and the variability. Central tendency describes where the distribution (that is, the collection of measurements) is located. To say that the measurements are located in the vicinity of, say, the value 15 is to say something about the central tendency. On the other hand, the variability describes how dispersed the data are. If we say that most of the data are within 5 units of the center of the distribution we are saying something about the variability of the distribution.

The following histograms (Figure 10.3) illustrate the distinction between central tendency and variability.

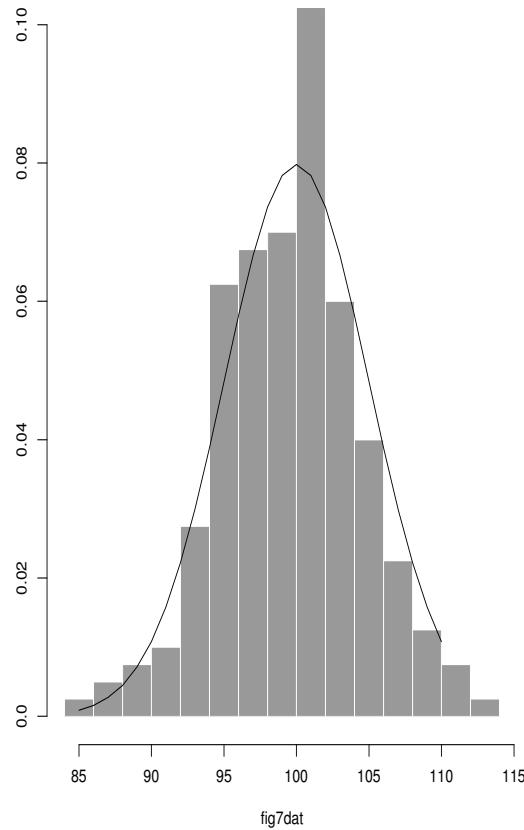


Figure 10.2: Histogram with density function superimposed, Section 10.2.

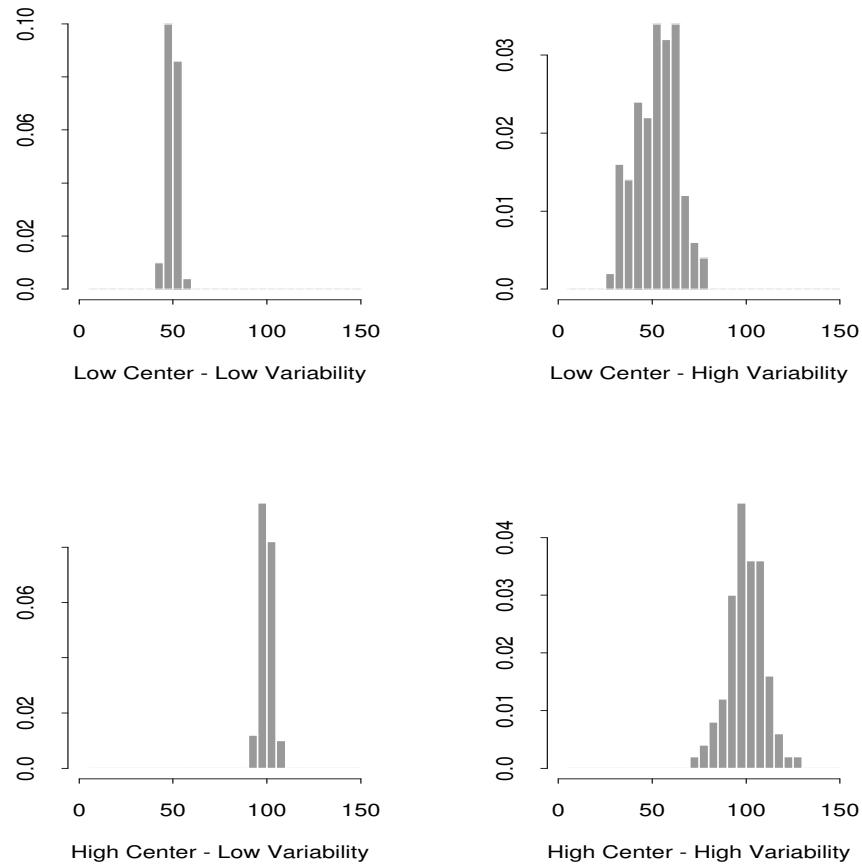


Figure 10.3: Histograms illustrating variability in central tendency and variability, Section 10.3.

The same trends can be illustrated using density curves (Figure 10.4).

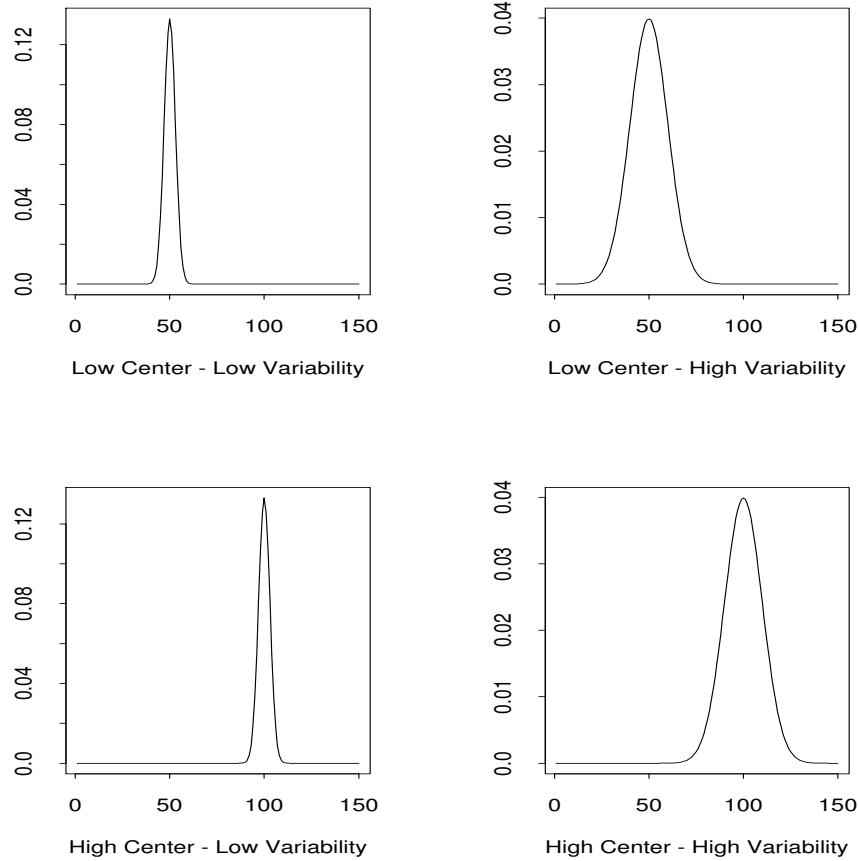


Figure 10.4: Densities illustrating variability in central tendency and variability, Section 10.3.

We saw in the last section a number of ways to numerically express central tendency using the mean, median and trimmed mean. The interquartile range (IQR) can be used as a measure of variability. But the most common measure of variability in data is the *sample variance*.

$$X_1, X_2, \dots, X_n$$

The formula for the (sample) variance is

$$S_n^2 = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{n - 1}$$

where  $\bar{X}_n$  is the sample mean (note that we have now added the subscript  $n$  to the notation  $\bar{X}$ ). Intuitively, to calculate  $S_n^2$  we take an observation  $X_i$

and subtract the mean  $\bar{X}_n$ , giving the distance between  $X_i$  and  $\bar{X}_n$ . Note, however, that this can be positive or negative depending on whether  $X_i$  is above or below  $\bar{X}_n$ . If we square the distance, then the result will always be positive. We do this for each observation  $X_i$  then sum the results. Finally we divide by  $n - 1$  (we'll see why we don't divide by  $n$  after some probability theory).

**Example 10.1** Suppose we want to calculate the sample variance of the following data.

$$50.3 \quad 52.5 \quad 58.6 \quad 62.9 \quad 64.0$$

We first need to calculate the mean.

$$\begin{aligned}\bar{X}_5 &= \frac{50.3 + 52.5 + 58.6 + 62.9 + 64.0}{5} \\ &= 57.66.\end{aligned}$$

The calculations needed are done in the following table

$i$	$X_i$	$X_i - \bar{X}_5$	$(X_i - \bar{X}_5)^2$
1	50.3	-7.36	54.17
2	52.5	-5.16	26.63
3	58.6	0.94	0.89
4	62.9	5.24	27.46
5	64.0	6.34	40.20
total			149.35

To complete the calculation we have from the table

$$\sum_{i=1}^5 (X_i - \bar{X}_5)^2 = 149.35$$

so that, with  $n = 5$

$$\begin{aligned}S_n^2 &= \frac{149.35}{5 - 1} \\ &= 37.3\end{aligned}$$

which gives the variance. ■

There are various ways to calculate  $S_n^2$ , given by

$$S_n^2 = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{n - 1} \tag{10.1}$$

$$= \frac{\sum_{i=1}^n X_i^2 - n\bar{X}_n^2}{n - 1} \tag{10.2}$$

$$= \frac{\sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n}}{n - 1} \tag{10.3}$$

We already used formula (10.1). Formulae (10.2) and (10.3) tend to be simpler to use. To see this we'll repeat the last calculation using (10.2).

**Example 10.2** To repeat the calculation of the previous example using formula (10.2), we can use much the same technique, except that a step is eliminated whereby we subtract  $\bar{X}_n$  from  $X_i$ .

$i$	$X_i$	$X_i^2$
1	50.3	2530.09
2	52.5	2756.25
3	58.6	3433.96
4	62.9	3956.41
5	64.0	4096.00
total		16772.71

From the table we have

$$\sum_{i=1}^5 X_i^2 = 16772.71.$$

We already know

$$\bar{X}_5 = 57.66$$

and that  $n = 5$ . Then, substituting into formula (2) gives

$$\begin{aligned} S_n^2 &= \frac{16772.71 - 5 \times 57.66^2}{5 - 1} \\ &= 37.3 \end{aligned}$$

which is the same value previously obtained. ■

The variance has an obvious and direct effect on the distribution. Figure 10.5 illustrates this. It consists of histograms from 4 data sets with mean 0 and varying variances. Note that as the variance increases, so does the degree to which the data are "spread out" (all histograms are drawn on the same range for comparison).

An important derived measure is the *sample standard deviation*. This is simply the square root of the variance

$$S_n = \sqrt{S_n^2}$$

As will be seen, it is often more natural to work with the standard deviation than with the variance. One reason for this is that it is necessarily in the same units as the original measurements.

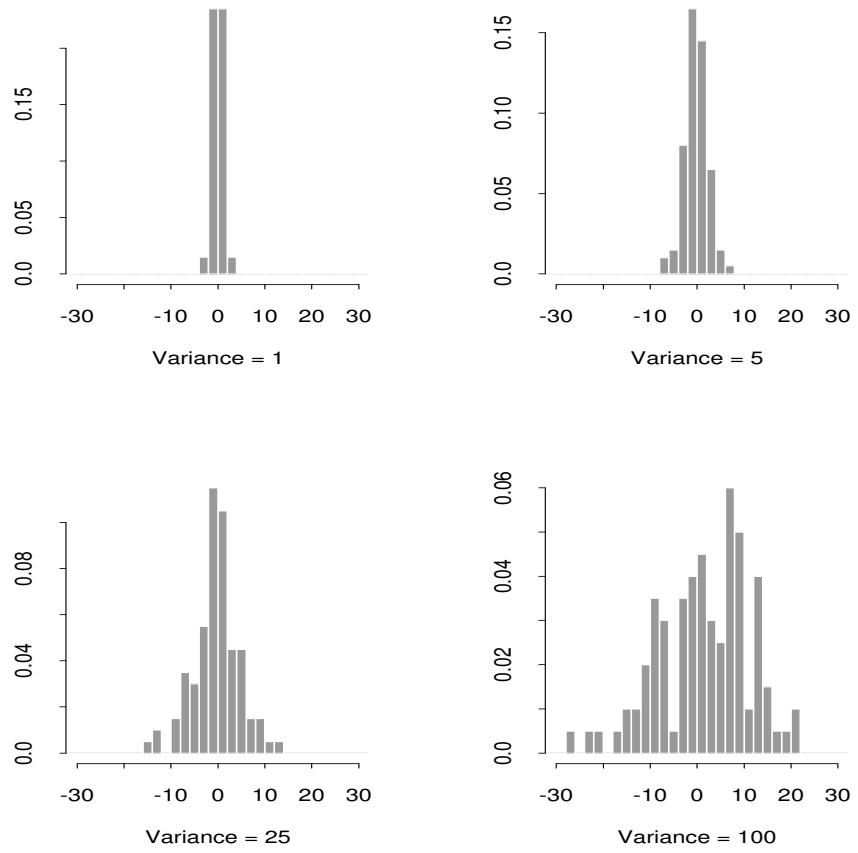


Figure 10.5: Effect of variance on a density

### 10.3.1 Updating Formula

It might come as a relief to know that if a new data point is added to a sample of size  $n$ , we don't need to recalculate  $\bar{X}_{n+1}$  and  $S_{n+1}^2$  (the sample mean and variance for the, now,  $n+1$  measurements). The advantage of the subscript  $n$  is now apparent, since we can now refer to the previous mean and variance  $\bar{X}_n$  and  $S_n^2$ . What we would like to be able to do is calculate  $\bar{X}_{n+1}$  and  $S_{n+1}^2$  directly from  $\bar{X}_n$  and  $S_n^2$  and the new measurement  $X_{n+1}$ . The sample mean is simply

$$\begin{aligned}\bar{X}_{n+1} &= \frac{\sum_{i=1}^{n+1} X_i}{n+1} \\ &= \frac{\sum_{i=1}^n X_i}{n+1} + \frac{X_{n+1}}{n+1} \\ &= \frac{n}{n+1} \bar{X}_n + \frac{X_{n+1}}{n+1}\end{aligned}\tag{10.4}$$

The sample variance is somewhat more complicated, but the principle is much the same, and yields an updating formula:

$$S_{n+1}^2 = \frac{n-1}{n} S_n^2 + \frac{(X_{n+1} - \bar{X}_n)^2}{n+1}.\tag{10.5}$$

We may adopt the convention that  $S_1^2 = 0$ , since only one measurement cannot be said to have any variation.

**Example 10.3** We may apply the formula (10.4) and (10.5) to the data of Example 10.1. In this case,  $\bar{X}_1$  and  $S_1^2$  are the mean and variance of the sample  $X_1 = 50.2$ ,  $\bar{X}_2$  and  $S_2^2$  are the mean and variance of the sample  $X_1 = 50.2, X_2 = 52.5$ , and so on. We have already argued that  $S_1^2 = 0$ . Also, we must have  $\bar{X}_1 = X_1 = 50.2$ , so the updating process can start at  $n = 2$ . We get:

$i$	$X_i$	$\bar{X}_i$	$S_i^2$
1	50.3	50.3	0
2	52.5	51.4	2.42
3	58.6	53.8	18.49
4	62.9	56.075	33.03
5	64	57.66	37.33

The previously obtained values of  $\bar{X}_5 = 57.66$  and  $S_5^2 = 37.3$ , can be found in the final row. ■

The formula (10.2) and (10.3) should be avoided for all but the smallest samples, since the magnitude of the sums can grow very quickly, causing significant rounding or magnitude errors. The use of updating formulae are to be preferred as general computing methods.

### 10.3.2 Variance in R

Sample variance is calculated by the function `var()`:

```
> x = rnorm(n = 200, mean = 0, sd = 4)
> var(x)
[1] 16.23771
> sqrt(var(x))
[1] 4.029604
> sd(x)
[1] 4.029604
```

The function `sd()` simply gives the standard deviation as the square root of `var()`.

## 10.4 Coefficient of Variation

One quantity which is sometimes used in statistical analysis is the *coefficient of variation*

$$CV = \frac{S_n}{\bar{X}_n},$$

that is, the ratio of the standard deviation to the mean (some conventions multiply this quantity by 100%). If  $\bar{X}_n$  is used as an estimate of some quantity  $\mu$ , the error is generally proportional to  $S_n$ , which is in the same units as  $\bar{X}_n$  (we will see why in later chapters). Therefore,  $CV$  can serve as an index of the accuracy available for a statistical estimation problem ( $\mu$  is assumed to be positive).

## 10.5 Symmetry and Skewness

One more important property of distributions is the degree of *skewness*. We say that a distribution is *symmetric* if the shape of the distribution is symmetric about the center. The degree of skewness of a distribution refers to the degree to which a distribution departs from symmetry.

Note the three distributions below (Figure 10.6). In the symmetric distribution, the center is located at 2. The distribution above 2 is approximately a mirror image of the distribution below 2. For the skewed distributions this is not the case. The right skewed distribution has a long right tail and a short left tail. The reverse is true for the left skewed distribution. Note that each of the three distributions has the same mean and variance (2 and 2).

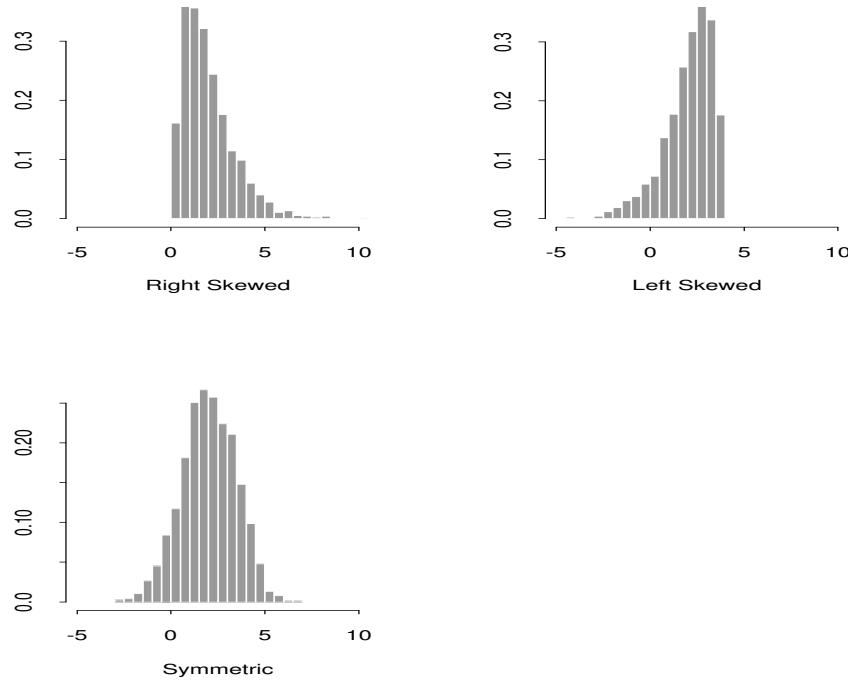


Figure 10.6: Examples of skewness

Whether or not a distribution is skewed can have important implications for analysis. Usually, statisticians find it easier to work with data which has a symmetric distribution. There should therefore be some numerical checks for skewness. If skewness is present, it may be desirable to transform the data in such a way as to reduce the skewness.

If a distribution has a long right tail, that means it has a relatively small number of very high measurements which have a disproportionate effect on the mean, since there are no correspondingly low measurements at the lower tail. However this does not affect the median. So for a right skewed distribution we would expect the mean to be higher than the median. Similarly, for a left skewed distribution we would expect the mean to be lower than the median. For a symmetric distribution we would expect the mean to be approximately equal to the median. The following table gives the mean and medians for the above histograms

Distribution	Mean	Median
Right Skewed	1.98	1.68
Left Skewed	2.04	2.31
Symmetric	2.02	2.02

A direct measure of skewness is given by the following formula

$$SK = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^3}{\left[ \sum_{i=1}^n (X_i - \bar{X}_n)^2 \right]^{3/2}}$$

This value will be positive for right skewed distributions, negative for left skewed distributions and approximately 0 for symmetric distributions. Note that  $SK$  has been standardized to be *unitless*.

It's worth noting that skewness tends to occur with certain kinds of data, particularly lifetimes and incomes. In both cases there exists a natural limit in the lower tail below which no observation would be found, but no such limit exists in the higher tail.

## 10.6 The Empirical Rule

As will be seen, normality of data is a crucial, and often over looked, assumption for many statistical procedures. It is therefore important to be able to test this assumption (and similar assumptions for other distributions).

One simple method for assessing normality is the *empirical rule*. Under the normal distribution the proportion of observations located within some number of standard deviations  $\sigma$  of the mean  $\mu$  is fixed for any  $\mu$  and  $\sigma$ . We may substitute sample estimates  $\bar{x}$  and  $S^2$  for  $\mu$  and  $\sigma$  and then examine these proportions.

Mathematically, the empirical rule is based on probabilities of the form

$$\begin{aligned} P(\mu - K\sigma \leq X \leq \mu + K\sigma) &= P\left(-K \leq \frac{X - \mu}{\sigma} \leq K\right) \\ &= P(-K \leq Z \leq K) \\ &= 1 - 2F_Z(-K). \end{aligned}$$

for any  $K > 0$ . The empirical rule holds for any mean and standard deviation  $\mu, \sigma$ . From the tables we get  $F_Z(-3) = 0.0013$ ,  $F_Z(-2) = 0.0228$  and  $F_Z(-1) = 0.1587$ , which gives

$$\begin{aligned} P(\mu - 1 \times \sigma \leq X \leq \mu + 1 \times \sigma) &= 1 - 2F_Z(-1) \approx 1 - 2 \times 0.1587 = 0.6826 \\ P(\mu - 2 \times \sigma \leq X \leq \mu + 2 \times \sigma) &= 1 - 2F_Z(-2) \approx 1 - 2 \times 0.0228 = 0.9544 \\ P(\mu - 3 \times \sigma \leq X \leq \mu + 3 \times \sigma) &= 1 - 2F_Z(-3) \approx 1 - 2 \times 0.0013 = 0.9974 \end{aligned}$$

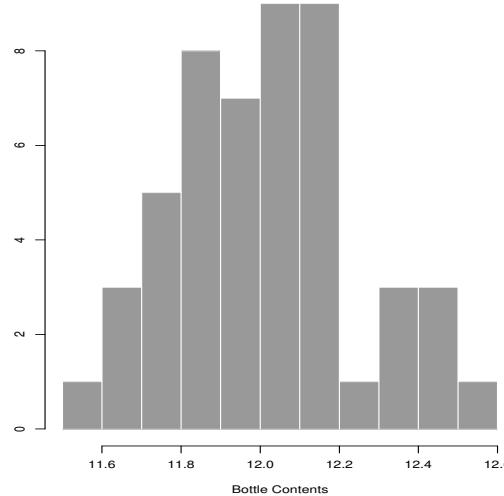


Figure 10.7: Histogram of data from Example 10.4.

from which the empirical rule is derived after rounding. When the data's distribution is symmetric and "bell shaped", that is, has one distinct frequency peak at its center, then the *empirical rule* will usually hold approximately.

**Definition 10.1** The *empirical rule* holds for a set of measurements if the following statements are approximately true:

- 68% of the data are within 1 standard deviation of the mean.
- 95% of the data are within 2 standard deviations of the mean.
- 99.7% (or most) of the data are within 3 standard deviations of the mean. ■

**Example 10.4** The following data represent the measured contents of 50 bottles of soda.

12.335	12.151	12.187	12.520	12.185	11.629	12.410	12.111
11.717	12.082	11.988	12.100	11.912	11.956	12.166	11.584
12.491	12.080	11.846	11.786	12.108	11.900	12.497	11.929
12.001	12.240	11.853	11.923	11.889	12.083	11.743	11.990
12.339	11.655	11.853	12.057	12.018	12.035	11.748	11.611
12.101	11.919	11.848	11.704	12.335	12.103	11.856	11.886
12.130	12.048						

A histogram of the data (Figure 10.7) suggests a bell shape. The tail is roughly equal in length on both sides.

We find that the mean and standard deviation are

$$\begin{aligned}\bar{X} &= 12.013 \\ S_n &= 0.232.\end{aligned}$$

The following table shows us how closely the empirical rule applies

$K$	$\bar{X} - KS_n$	$\bar{X} + KS_n$	% within Range (Theoretical)	% within Range (Actual)
1	11.781	12.245	68%	70%
2	11.549	12.477	95%	94%
3	11.317	12.709	99.7%	100%

The data can be seen to conform very closely to the empirical rule. ■

## 10.7 Quantile Plots

The empirical rule provides a simple method for assessing the normality of a set of data. The *quantile plot* provides a more refined method for assessing distributional assumptions for the normal and other distributions.

Suppose we have a set of observations  $X_1, \dots, X_n$ , and we wish to know they conform to a distribution with CDF  $F_X$ .

Recall from Section 8.6 that sample quantiles can be constructed from the order statistics

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}.$$

For a list of probabilities  $p_1, \dots, p_m$  we may compare sample quantiles

$$\hat{Q}(p_1), \dots, \hat{Q}(p_m)$$

to the theoretical quantiles

$$Q(p_1), \dots, Q(p_m)$$

obtained directly from the quantile function  $Q$ , which follows from  $F_X$  (see Definition 4.4). A quantile plot is a scatter plot of the sample against the theoretical quantiles. If the plot lies approximately on the identity then the data conforms to distribution  $F_X$ .

Quantile plots are usually constructed by taking the order statistics themselves as the sample quantiles

$$\hat{Q}(p_1) = X_{(1)}, \dots, \hat{Q}(p_n) = X_{(n)}.$$

We next need to consider the appropriate choices to  $p_1, \dots, p_n$ . In Section 8.6 we considered the problem of constructing a sample  $p$ -quantile. Here, we need to decide for which  $p$   $X$  is a  $p$ -quantile. It is tempting to set

$$p_k = k/n \text{ so that } X_{(k)} = \hat{Q}(k/n) \approx Q(k/n).$$

However, this would mean  $p_n = n/n = 1$ . However  $Q(1)$  is the upper bound of the support of  $X$ , and we have no reason to think that  $X_{(n)}$  estimates this number, even though it is the maximum observation. Note that for a normal distribution we have  $Q(1) = \infty$ .

There is no single answer to this problem. Different criterion will give different solutions (see the discussion in Section 8.6 on this point). For the standard method in R for creating quantile plots the probability points used are given by the formula

$$p_k = \frac{k - a}{n + (1 - a) - a}, \quad \text{where } a = 3/8 \text{ for } n \leq 10, \quad \text{and } a = 1/2 \text{ for } n > 10, \quad (10.6)$$

then the quantile plot is a scatter plot of the pairs  $(Q(p_i), X_{(i)})$  for  $i = 1, \dots, n$ .

However, we note that when we test the assumption of normality, we need not be concerned with the actual value of the mean and variance. However, to construct the theoretical quantiles  $Q(p_i)$  we do need to specify an exact distribution. Recall from Section 4.12.4 that if  $Z_\alpha$  is the  $\alpha$ -quantile of distribution  $N(0, 1)$ , then

$$X_\alpha = \mu + \sigma Z_\alpha$$

is the  $\alpha$ -quantile of distribution  $N(\mu, \sigma^2)$ . Therefore, if we use a formula such as 10.6 to generate probability points, calculate theoretical quantiles  $Q(p_i)$  from the *standard* normal distribution  $N(0, 1)$ , then plot pairs  $(Q(p_i), X_{(i)})$ , then the scatter plot should be approximately a straight line if the data is from *any* normal distribution  $N(\mu, \sigma^2)$ .

In R normal quantile plots are produced by the function `qqnorm()`. The function `qqline()` superimposes a straight line to assist in the comparison. If the data is contained in vector `x` the sequence of commands is typically:

```
> qqnorm(x)
> qqline(x)
```

See Section 10.8.3 for more examples using these functions.

**Example 10.5** Figure 10.8 shows normal quantile plots for samples of size  $n = 200$  from a uniform distribution on  $[0, 1]$ , and exponential distribution of rate 1, and a standard normal distribution. A straight line is included

to clarify the plot. For the normal distribution, the quantile plot conforms closely to the straight line predicted by theory, while for the nonnormal samples, the deviation from linearity is evident. ■

## 10.8 Transformations

A statistical analysis often involves some transformation of data, done before the analysis is carried out. There are many reasons for this. There may be a practical necessity to change measurement units. Such a transformation will generally not affect the type of analysis done. In the other hand, the data analyst will sometimes transform data for the purposes of making some analytical method more simple, accurate or effective.

### 10.8.1 Linear Transformations

Linear transformations are usually performed on data to convert units or to simplify data coding. A linear transformation takes the form

$$Y = aX + b$$

where  $X$  is the original data,  $Y$  is the new transformed data, and  $a$  and  $b$  are two constants which define the transformation. Below are some common transformations

Convert from ( $X$ )	Convert to ( $Y$ )	$a$	$b$	Formula
Celsius	Fahrenheit	1.8	32.0	$Y = 1.8X + 32.0$
Fahrenheit	Celsius	0.556	-17.792	$Y = 0.556X - 17.792$
Meters	Feet	3.2808	0	$Y = 3.2808X$
Feet	Meters	0.3048	0	$Y = 0.3048X$
Kilograms	Pounds	2.2046	0	$Y = 2.2046X$
Pounds	Kilograms	0.45359	0	$Y = 0.45359X$

For most unit conversions  $b$  is set to 0. One of the advantages of the linear transformation is that many of the statistical summaries do not have to be explicitly recalculated when the data are subjected to one. If we have data

$$X_1, X_2, \dots, X_N$$

and we create a new data set

$$Y_1, Y_2, \dots, Y_N$$

by applying a linear transformation

$$Y_i = aX_i + b$$

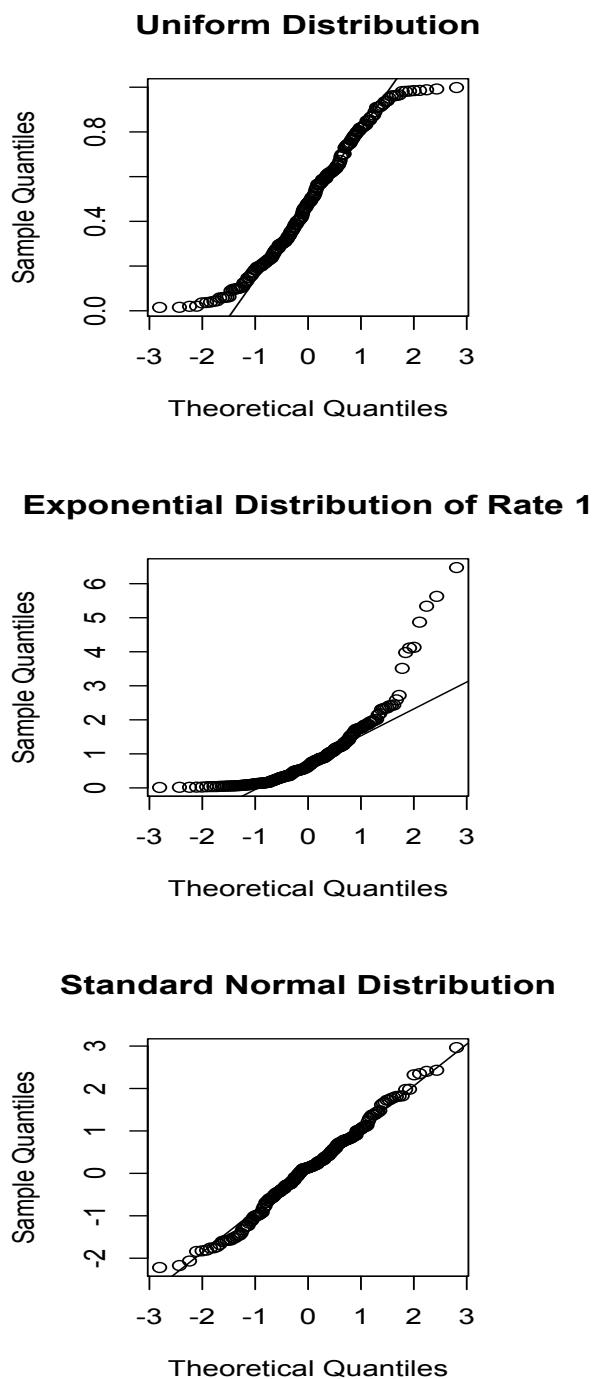


Figure 10.8: Normal quantile plots for samples of size  $n = 200$  from a uniform distribution on  $[0, 1]$ , and exponential distribution of rate 1, and a standard normal distribution.

to each value  $X_i$ , for some fixed values  $a$  and  $b$  then the following table gives conversion methods for each of the statistical summaries previously discussed.

Statistic	Old Value ( $X_i$ )	New Value ( $Y_i$ )
Mean	$\bar{X}$	$a\bar{X} + b$
Median	$\tilde{X}$	$a\tilde{X} + b$
Trimmed Mean	$\bar{X}_{K\%}$	$a\bar{X}_{K\%} + b$
Variance	$S_n^2$	$a^2 S_n^2$
Standard Deviation	$S_n$	$ a S_n$
Inter Quartile Range	$IQR$	$ a IQR$
Lower Quartile	$Q_1$	$aQ_1 + b$ if $a > 0$ $aQ_3 + b$ if $a < 0$
Upper Quartile	$Q_3$	$aQ_3 + b$ if $a > 0$ $aQ_1 + b$ if $a < 0$
Quantile	$X_{K\%}$	$aX_{K\%} + b$ if $a > 0$ $aX_{(100-K)\%} + b$ if $a < 0$

### 10.8.2 Transformations to Reduce Skewness

The empirical rule depends on the data being approximately symmetric. In fact, many of the procedures discussed later assume that the data is symmetric. If this is not the case, than the conclusions reached may not be accurate.

As we have already seen, there are many types of data which are not naturally symmetric. This tends to be the case with data which represents dollar amounts, as well as data which measures age or survival time. It is sometimes best, therefore, to subject the data to a transformation that makes the resulting data set more symmetric. Analysis is then preformed on the transformed data.

We need to find an appropriate function  $f(x)$  with which to transform a data set

$$X_1, X_2, \dots, X_N$$

to the data set

$$Y_1, Y_2, \dots, Y_N$$

by applying the transformation

$$Y_i = f(X_i).$$

We will insist that  $f(x)$  be an increasing function, so that if

$$X_i > X_j$$

we will also have

$$Y_i > Y_j.$$

If the data is skewed rightward, we also want a function that tends to reduce larger values in proportion to the smaller ones. This can be achieved by using a function which is increasing, but whose slope is decreasing. Common choices are the logarithm function

$$f(x) = \log(x)$$

or the square root function

$$f(x) = \sqrt{x}.$$

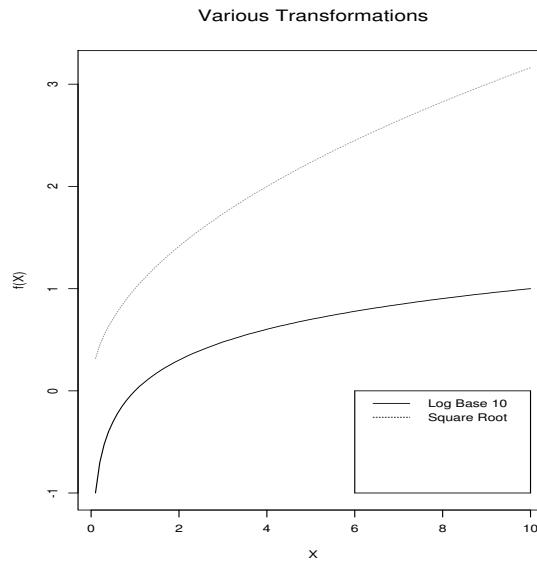


Figure 10.9: Commonly used transformations to eliminate right skewness, Section 10.8.2.

**Example 10.6** The General Social Survey is conducted (almost) annually by the National Opinion Research Center at the University of Michigan. This survey collects data from a sample of Americans on many demographic and lifestyle characteristics.

We present here a histogram of data from the 1993 survey giving the age at which respondents first married. There were 1202 responses to this variable in the survey. There is definite right skewness, so we apply the two transformations introduced above.

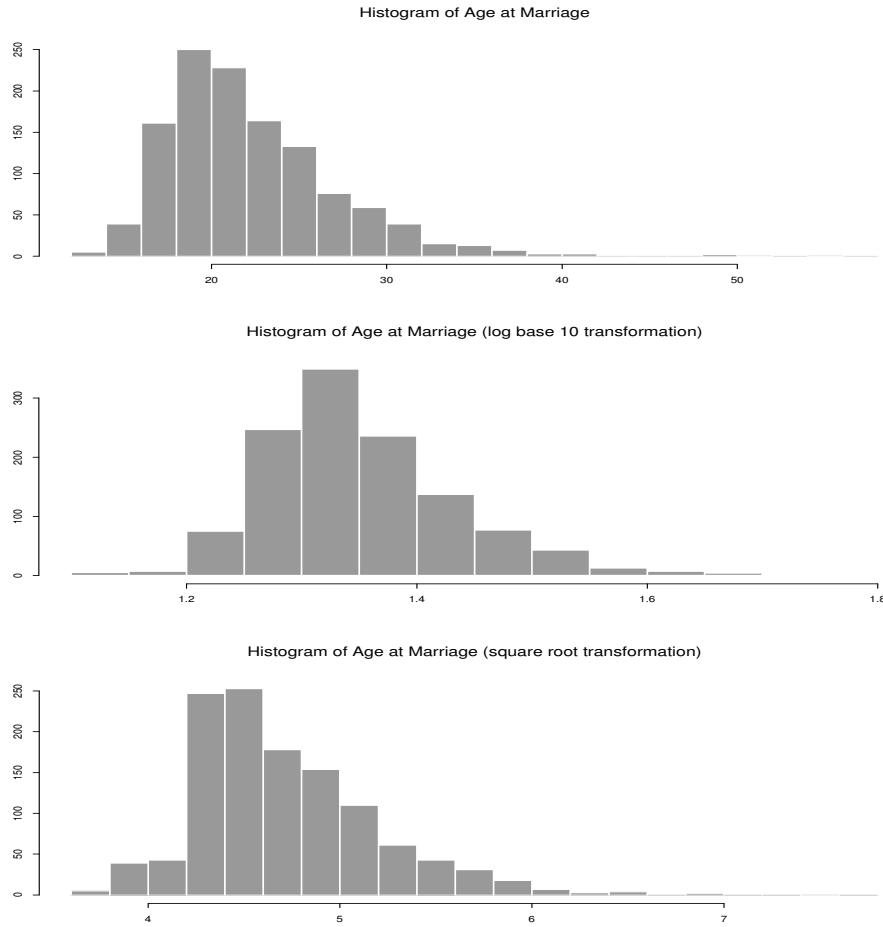


Figure 10.10: Histograms marriage age data with transformations for Example 10.6.

Both transformations reduce skewness, but the log (base 10) transformation appears to do so to a larger degree, although even with this transformation skewness still appears to be present (see Figure 10.10). As a check, we apply the empirical rule to the original data and to the two transformed data sets.

	$K$	$\bar{X} - KS_n$	$\bar{X} + KS_n$	% within Range (Theoretical)	% within Range (Actual)
Untransformed	1	17.76	27.82	68	78.37
	2	12.73	32.86	95	96.01
	3	7.69	37.89	99.7	98.59
Log (base 10)	1	1.26	1.44	68.0	68.55
	2	1.17	1.52	95.0	96.42
	3	1.09	1.61	99.7	99.17
Square	1	4.25	5.25	68.0	68.55
Root	2	3.75	5.74	95.0	95.59
Transformed	3	3.25	6.243	99.7	98.92

The original data departs significantly from the empirical rule, especially in the  $\pm 1$  standard deviation frequency. The two transformed data sets conform well to the empirical rule, despite the fact that some skewness remains after the transformation. ■

**Example 10.7** We continue with Example 10.6. A normal probability plot is given in Figure 10.11 for the original marriage age data, and for the two transformations (see Section 10.7). The original data gives a decidedly non-linear curve, as we might expect from the large amount of skewness present. The two transformed data sets give a curve which is somewhat more linear, although some systematic curvature is still present. By visual inspection, the log transformed data comes closest to being normally distributed. ■

### 10.8.3 Box-Cox Power Transformations

Many laboratory measurements are conventionally log-transformed for statistical analysis due to the technical nature of the assay process. The underlying assumption is that the untransformed assay is log-normally distributed (See, for example, Horne *et al.* (2001)). Somewhat more flexibility is provided by the family of *Box-Cox power transformations*, indexed by parameter  $\lambda$ . For positive value  $y$  the (one parameter) transformed value  $y_\lambda$  is defined as

$$y_\lambda = \begin{cases} \frac{y^\lambda - 1}{\lambda} & ; \lambda \neq 0 \\ \log(y) & ; \lambda = 0 \end{cases}$$

where the piecewise definition is justified by the fact that for any fixed  $y > 0$  it can be shown that  $(y^\lambda - 1)/\lambda$  converges to the natural logarithm  $\ln(y)$  as  $\lambda$  approaches 0. When the objective is to regress  $y$  onto predictor variables  $X$  the appropriate value of  $\lambda$  can be estimated by fitting models  $z_\lambda = \beta X$ , selecting  $\lambda$  which minimizes the residual sum of squares (RSS), where  $z_\lambda$  is the standardized power transformation

$$z_\lambda = \frac{y_\lambda}{\text{GM}(y)^{\lambda-1}}, \quad (10.7)$$

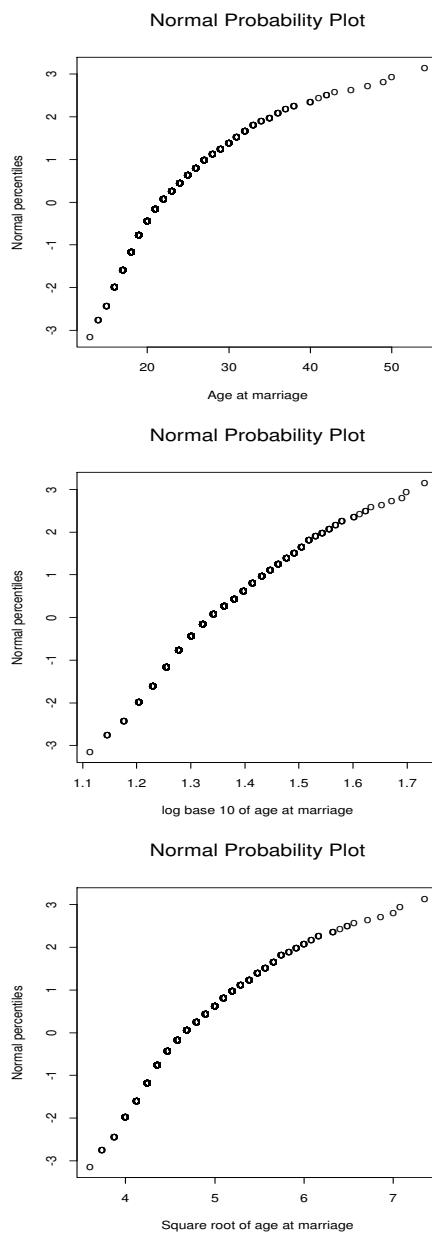


Figure 10.11: Normal quantile for marriage age data with transformations

and  $\text{GM}(y)$  is the geometric mean of the observations  $y$ . In this way, the suitability of the log-transformation can be assessed, and an alternative transformation employed when appropriate.

In R we can used the `boxcox()` function in library MASS

```
> boxcox
Error: object 'boxcox' not found
> library(MASS)
Warning message:
package 'MASS' was built under R version 3.0.2
> boxcox
function (object, ...)
UseMethod("boxcox")
<bytecode: 0x100c147b8>
<environment: namespace:MASS>
>
```

The function returns a list consisting of elements labeled  $x$  (values of  $\lambda$ ) and  $y$  (the likelihood, inversely related to the RSS). The Box-Cox transformation is taken to be the value of  $\lambda$  which maximizes the likelihood (equivalently, minimizes RSS). By default, `boxcox()` plots the likelihood against  $\lambda$ , superimposing a 95% confidence interval. If the function is copied into an object the maximum-likelihood value of  $\lambda$  can be extracted as follows:

```
> y = exp(rnorm(100))
> bc.obj = boxcox(exp(y)~1)
> lambda = bc.obj$x[bc.obj$y == max(bc.obj$y)]
> lambda
[1] -0.4646465
```

Figure 10.12 shows the resulting plot.

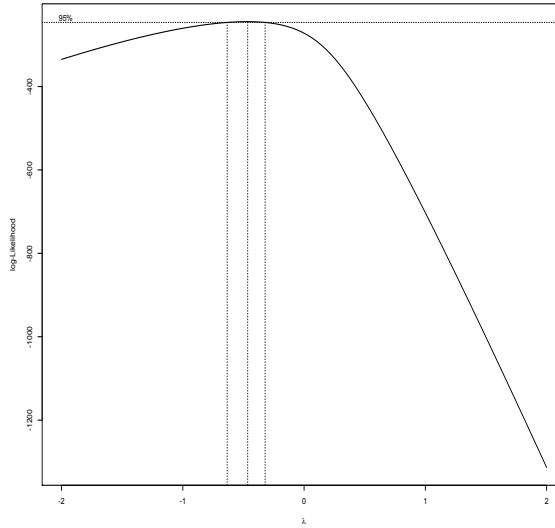


Figure 10.12: Example of likelihood plot produced by `boxcox()` function.

We may wish to write our own function which returns  $\lambda$  directly, as well as the transformed data. Note that the `lambda` option appearing in the `boxcox()` function increases the resolution of the plot.

```
boxcox.wrapper = function(y, pflag=F) {

  bc.obj = boxcox(y~1, , lambda = seq(-4, 4, 1/100), plotit=pflag)
  attributes(bc.obj)
  bc.obj$x[1:10]
  bc.obj$y[1:10]
  lambda = bc.obj$x[bc.obj$y == max(bc.obj$y)]

  if (lambda == 0 ) {
    ylambda = log(y)
  } else
  {
    ylambda = (y^lambda - 1)/lambda
  }
  return(list(ylambda = ylambda, lambda=lambda))
}
```

We can examine the Box-Cox transformation using the following script:

```
par(mfrow=c(3,4),cex=1.0)
```

```

f1 = function(n) 1/rchisq(n,df=4)
dist.list = c(rnorm, rexp, f1)
dist.lab = c('normal', 'exponential', 'reciprocal chi.sq 4df')
n = 100

for (i in 1:length(dist.list)) {

  fun = dist.list[[i]]
  lab = dist.lab[i]

  y = exp(fun(n))

  ylog = log(y)
  bc.obj = boxcox.wrapper(y,T)
  qqnorm(y, main = paste(dist.lab[i], '\n untransformed'))
  qqline(y)
  qqnorm(ylog,main = paste(dist.lab[i], '\n log transformed'))
  qqline(ylog)
  new.lambda = signif(bc.obj$lambda,3)
  qqnorm(bc.obj$ylambda,
         main = paste(dist.lab[i], '\n Box-Cox transformed'))
  qqline(bc.obj$ylambda)
  mtext(bquote(lambda == .(new.lambda)),side=3)
}

```

This script produces the plot in Figure 10.13. We make the following observations.

1. Normal quantile plots show the degree to which the transformation ‘normalizes’ the data. The log-normal, log-exponential and log-reciprocal  $\chi^2_4$ . are used as examples. As would be expected, for the normal data we get  $\lambda = 0.03$ , close to 0, meaning that the log-transform would be appropriate. For the remaining distributions we get  $\lambda = -0.66, -2.12$ . While the Box-Cox transform does not entirely succeed in ‘normalizing’ the data, it does succeed in reducing the skewness of both the untransformed and log-transformed data, leaving the data close to symmetrical, if not normal.
2. The `par(mfrow=c(3,4),cex=1.0)` command permits array-style plotting. The `cex` option controls display size of the plot features.
3. The object `dist.list` becomes a *list of functions*, not a vector. This makes it easier to automate studies of this kind. Note that a specialized function for the reciprocal- $\chi^2_4$  was created.

4. We can plot math symbols, including Greek notation. See `help(plotmath)`, `help(mtext)` and `help(bquote)`.
5. The '\n' characters in the plot titles force a line feed.

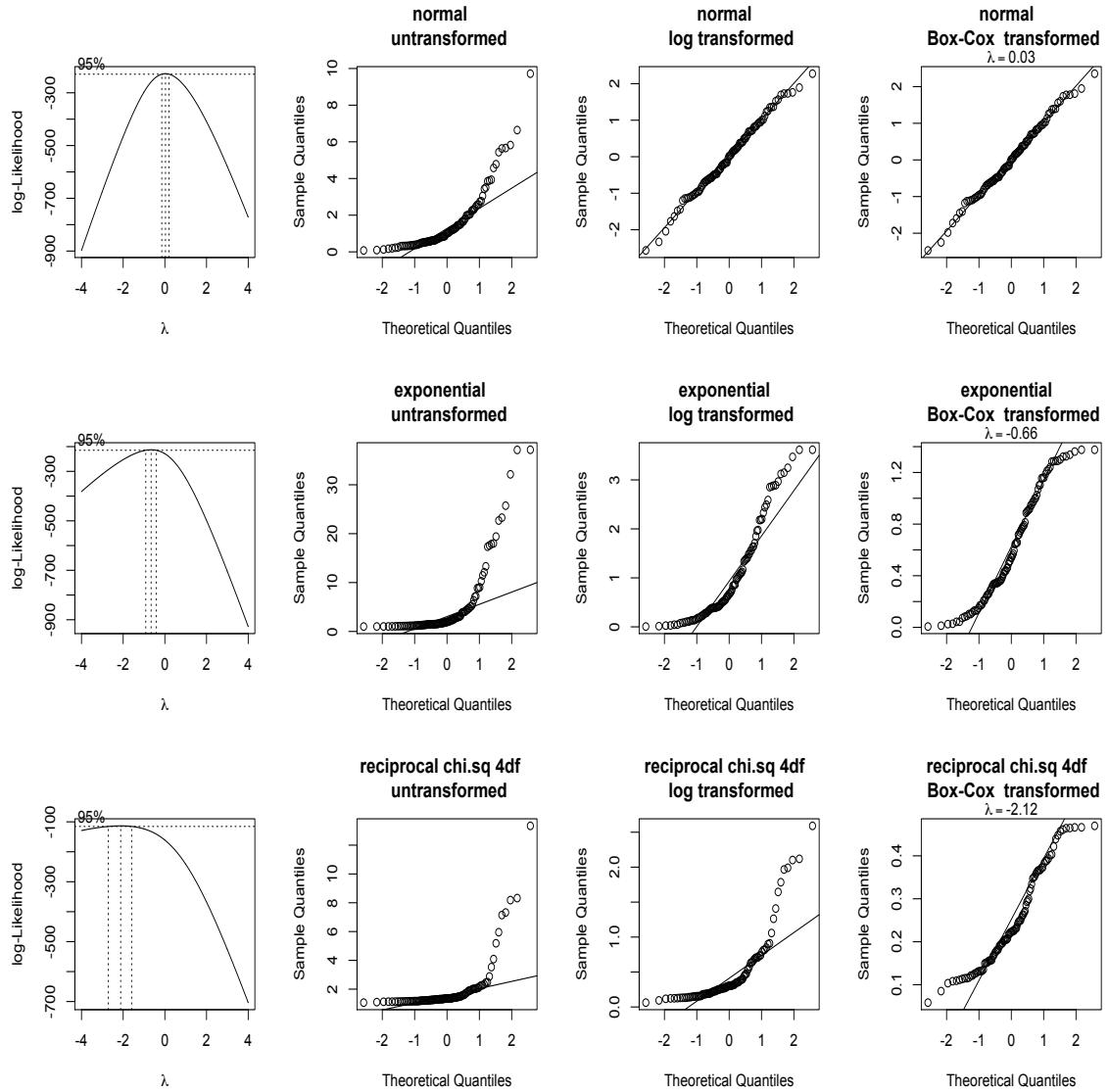


Figure 10.13: Assessment of Box-Cox transformations for log-normal, log-exponential and log-reciprocal  $\chi^2_4$ .

# Chapter 11

## Relationships Between Variables

So far we have concentrated our attention on one single variable. However, most statistical studies involve more than one variable, and a central objective of the analysis may be to establish the existence of relationships between them. Does income increase with years of education? Is the proportion of men who smoke higher than that of women?

To answer such questions, we usually have data consisting of multiple observations from the subjects of the study.

Name	Subject ID	Sex	Income	Years of Education	Smokes
Smith, John	10001	M	35,000	3	Heavy
Jones, Betty	10002	F	45,000	5	No
:	:	:	:	:	:

The manner in which such relationships are explored depends on the form of the data. We will look at three situations,

1. both variables are categorical,
2. one variable is categorical, the other is numerical,
3. both variables are numerical.

Some flexibility may be required with these definitions. If a variable is numerical, but the number of distinct values assumed in the data set is small (e.g. the number of cars owned by a family), it may be more sensible to treat it as an ordinal categorical variable.

## 11.1 Relationships Between Categorical Variables

The principal means of examining the relationship between two categorical variables is the *contingency table*. This is also known as a *cross tabulation* of two categorical variables. Formally, a contingency table gives the frequencies in a sample for every combination of values of the two variables.

Suppose we have the following data:

Subject ID	Sex	Smokes
1	M	Heavy
2	M	Heavy
3	M	Heavy
4	M	Light
5	M	Light
6	M	No
7	M	No
8	F	Heavy
9	F	Heavy
10	F	Heavy
11	F	Heavy
12	F	Light
13	F	Light
14	F	Light
15	F	No
16	F	No
17	F	No
18	F	No
19	F	No
20	F	No

In all, there are 6 combinations of categories:

1. Male heavy smoker
2. Male light smoker
3. Male nonsmoker
4. Female heavy smoker
5. Female light smoker
6. Female nonsmoker.

To represent the data in table form we designate one variable to be the *row variable* and the other to be the *column variable*. Here we'll use sex as the

row variable. We construct one row for each row variable value, and one column for each column variable value. Then the entry for each row and column combination is the number in the sample with that combination of values. For example, in the preceding data set there are 3 male heavy smokers and 6 female nonsmokers.

	Heavy Smoker	Light Smoker	Non Smoker	Total
Male	3	2	2	7
Female	4	3	6	13
Total	7	5	8	20

The contingency table is usually given with row and column totals. For example, we can see from the table that there are 13 females, and 5 light smokers in the sample.

Some care is needed in interpreting a contingency table. Although there are more female than male heavy smokers in the sample, we need to note that there are also more women than men in the sample. So we must determine if the fact that there are more female than male heavy smokers is simply due to the fact that there are more females in the sample. To do this, we may express the table frequencies as *row percentages*. That is, the frequency is given as a percentage of the total frequencies for that row.

	Heavy Smoker	Light Smoker	Non Smoker	Total
Male	3 (43%)	2 (29%)	2 (29%)	7 (100%)
Female	4 (31%)	3 (23%)	6 (46%)	13 (100%)
Total	7 (35%)	5 (25%)	8 (40%)	20 (100%)

We note from the table that 43% of males are heavy smokers but only 31% of females are heavy smokers. We can conclude that, at least in this sample, any given male is more likely to be a heavy smoker than any given female.

We may also calculate column percentages, which will be the frequencies expressed as a percentage of the total frequencies in a column.

	Heavy Smoker	Light Smoker	Non Smoker	Total
Male	3 (43%)	2 (40%)	2 (25%)	7 (35%)
Female	4 (57%)	3 (60%)	6 (75%)	13 (65%)
Total	7 (100%)	5 (100%)	8 (100%)	20 (100%)

Notice that although, any given male is more likely to be a heavy smoker than any given female, any given heavy smoker is more likely to be a female than a male (since there are more females).

**Example 11.1** The following contingency table appeared in the *American Journal of Epidemiology* (Martin & Bracken, 1987).

Marital status	Caffeine Consumption (mg/day)				Total
	0	1-150	151-300	> 300	
Married	652	1537	598	242	3029
Divorced, separated or widowed	36	46	38	21	141
Single	218	327	106	67	718
Total	906	1910	742	330	3888

First notice that the column variable is numerical, but has been collapsed into an ordinal categorical variable. This is often done to present results in a simple, clear manner.

We are interested in knowing whether or not there is any relationship between marital status and caffeine consumption. More specifically, is there any difference in caffeine consumption patterns between people of different marital status? Notice that there are large differences in the numbers of subjects in the different marital status categories. There are 3029 married subjects, but only 141 divorced, separated or widowed subjects, so interpreting the absolute frequencies can be highly misleading. We therefore give the same table using row percentages.

Marital status	Caffeine Consumption (mg/day)				Total
	0	1-150	151-300	> 300	
Married	22%	51%	20%	8%	100%
Divorced, separated or widowed	26%	33%	27%	15%	100%
Single	30%	46%	15%	9%	100%
Total	23%	49%	19%	8%	100%

Note that the percentage of divorced, separated or widowed subjects who are in the highest consumption category is 15, whereas for married subjects and single subjects it is 8 and 9 respectively. It therefore appears that divorced, separated or widowed subjects are most likely to be heavy

coffee drinkers. If we examine the lowest category of consumption we see that single subjects are most likely to abstain from caffeine altogether. We should point out that we must be able to show that this result would be unlikely to occur by chance. This will be covered in a later section. ■

## 11.2 Relationships Between One Categorical Variable and One Numerical Variable

We now suppose that we are interested in the relationship between two variables, one of which is categorical and the other numerical. The usual procedure is to split the observations into groups determined by the categorical variable. Then, a numerical or graphical summary of the numerical variable is presented for each group. If a graphical summary is chosen, the usual type of plot is a boxplot.

**Example 11.2** We return to the prostate cancer data examined in Example 9.3. Each subject was classified as a nonsmoker, light smoker or heavy smoker. The data were then split into the resulting three groups. The following table gives summary statistics of the survival times (in years) from diagnosis for each of the three groups.

	Nonsmoker	Light smoker	Heavy smoker
number of subjects	32	6	48
mean	12.38	10.83	9.25
standard deviation	3.49	6.27	4.34
1st quartile	10.75	6.00	5.75
median	12.50	10.00	11.00
3rd quartile	15.00	15.50	12.25
IQR	4.25	9.50	6.50

We have two measures of central tendency, the median and the mean. The mean decreases steadily from nonsmoker to heavy smoker. As for the median, the middle category has the lowest value, which is not what we would expect. Note, however that there are only 6 subjects in this category, which makes it difficult to assign any significance to this feature.

We also have two measures of variability, the standard deviation and the IQR. In both cases the order of variability is the same for each measure. An interesting feature is the steep drop in the 1st quartile from the nonsmoker category to both smoker categories. This suggests a relatively large number of smokers experience rapid deterioration of their condition, compared to the nonsmoking group.

Multiple boxplots give a clear picture of these trends. One boxplot is constructed for each group, and plotted using the same survival time axis to enable comparison. ■

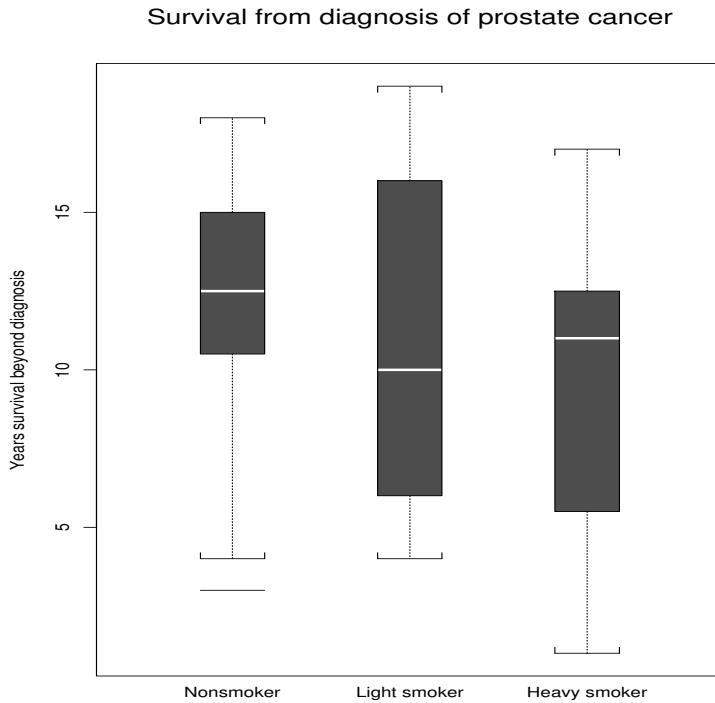


Figure 11.1: Boxplots from the prostate cancer study, Example 11.2.

### 11.3 Relationships Between Numerical Variables

There are numerous methods available for examining the relationship between two numerical variables. In fact, this topic is one of the principal areas of investigation among statistical methodologists. We present a summary of some of the most fundamental techniques.

#### 11.3.1 Scatter Plots

The *scatter plot* is a graphical technique used to visually inspect the relationship between two variables. Suppose we have the following 5 observations, each consisting of 2 measurements.

Observation	Variable 1	Variable 2
1	5.2	10.2
2	6.3	10.5
3	6.9	12.0
4	7.5	12.5
5	8.1	13.6

Suppose we have a graph with the horizontal axis representing variable 1 and the vertical axis representing variable 2. For each observation we place a point on the plot, with the horizontal coordinate given by variable 1 and the vertical coordinate given by variable 2.

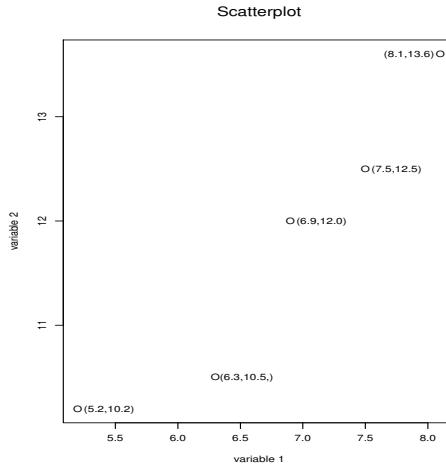


Figure 11.2: Example of scatter plot, Section 11.3.1.

The scatter plot suggests that there is a positive relationship between the two variables, in the sense that an increase in variable 1 means an increase in variable 2.

**Example 11.3** The scatter plot in Figure 11.3 represents 392 automobiles. For each car the miles per gallon rating and the horsepower rating were recorded. These two variables make up the scatter plot. As expected, the scatter plot reveals that the higher the horsepower, the lower the miles per gallon achieved. ■

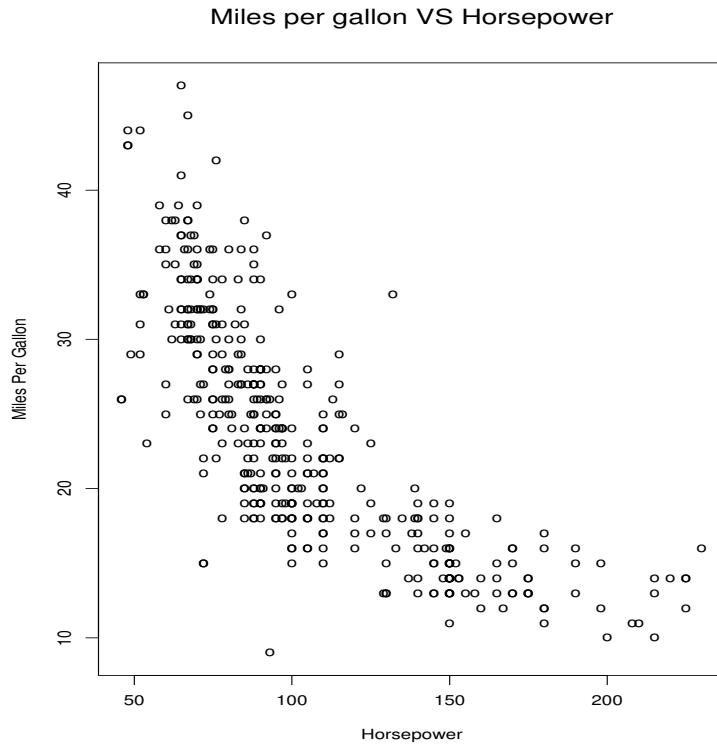


Figure 11.3: Scatter plot for automobile data, Example 11.3.

### 11.3.2 Correlation

It is possible to assign a numerical measure to the degree of association between two variables. Suppose measurements of two variables are given by

$$X_1, X_2, \dots, X_n$$

and

$$Y_1, Y_2, \dots, Y_n.$$

The *sample correlation coefficient* is defined by

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}.$$

The *sample covariance* is defined by

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}).$$

At this point it is worth introducing the idea of the *sum of squares* (notational conventions vary):

$$\begin{aligned} SS_X(n) &= \sum_{i=1}^n (X_i - \bar{X}_n)^2 \\ SS_Y(n) &= \sum_{i=1}^n (Y_i - \bar{Y}_n)^2 \\ SS_{XY}(n) &= \sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n), \end{aligned}$$

so we may write

$$r_n = \frac{SS_{XY}(n)}{\sqrt{SS_X(n)SS_Y(n)}}.$$

The sums of squares  $SS_X(n)$  and  $SS_Y(n)$  are directly related to the sample variances, for example

$$SS_X(n) = (n - 1)S_n^2.$$

The updating formulae of Section 3.3.1 may be used to calculate or update  $\bar{X}_n$ ,  $\bar{Y}_n$ ,  $SS_X(n)$ ,  $SS_Y(n)$  and  $SS_{XY}(n)$  separately, and so may be used to calculate or update  $r_n$  itself:

$$\begin{aligned} SS_X(n+1) &= SS_X(n) + \frac{n}{n+1}(X_{n+1} - \bar{X}_n)^2 \\ SS_Y(n+1) &= SS_Y(n) + \frac{n}{n+1}(Y_{n+1} - \bar{Y}_n)^2 \\ SS_{XY}(n+1) &= SS_{XY}(n) + \frac{n}{n+1}(X_{n+1} - \bar{X}_n)(Y_{n+1} - \bar{Y}_n), \end{aligned}$$

with  $SS_X(1) = SS_Y(1) = SS_{XY}(1) = 0$ .

**Example 11.4** We will use the data from the previous subsection to illustrate the calculation. After relabelling appropriately, the data is given by

$i$	$X_i$	$Y_i$
1	5.2	10.2
2	6.3	10.5
3	6.9	12.0
4	7.5	12.5
5	8.1	13.6

Using the formulae of the previous section we get

$$\begin{aligned}\bar{X} &= 6.8 \\ \bar{Y} &= 11.76 \\ \sum_{i=1}^n (X_i - \bar{X})^2 &= 5.00 \\ \sum_{i=1}^n (Y_i - \bar{Y})^2 &= 8.01\end{aligned}$$

To calculate the numerator of  $r$  we can use the following table.

$i$	$X_i$	$X_i - \bar{X}$	$Y_i$	$Y_i - \bar{Y}$	$(X_i - \bar{X})(Y_i - \bar{Y})$
1	5.2	-1.6	10.2	-1.56	2.496
2	6.3	-0.5	10.5	-1.26	0.630
3	6.9	-0.1	12.0	0.24	-0.024
4	7.5	0.7	12.5	0.74	0.518
5	8.1	1.3	13.6	1.84	2.392
Total					6.06

The correlation is then given by

$$\begin{aligned}r &= \frac{6.06}{\sqrt{5.00 \times 8.01}} \\ &= 0.957\end{aligned}$$

■

**Example 11.5** The correlation of the miles per gallon and horsepower measurements of the automobile example was found to be -0.77. ■

It is a mathematical fact that  $r$  is always in between -1 and +1. When  $r$  is negative the two variables are negatively correlated, meaning that when one increases the other tends to decrease. When  $r$  is positive, the two variables are positively correlated, meaning that the two variables increase and decrease together. When  $r$  is near zero, then the two variables are uncorrelated, meaning that the value of one has little information about the value of the other. The scatter plot of Figure 11.4 are given with their correlations.

### 11.3.3 Correlations and Covariances in R

Covariance and correlation in R are given by the functions `cov()` and `cor()`:

```
> x1 = rnorm(200,0,1)
> x2 = rnorm(200,0,1)
> cov(x1,x2)
```

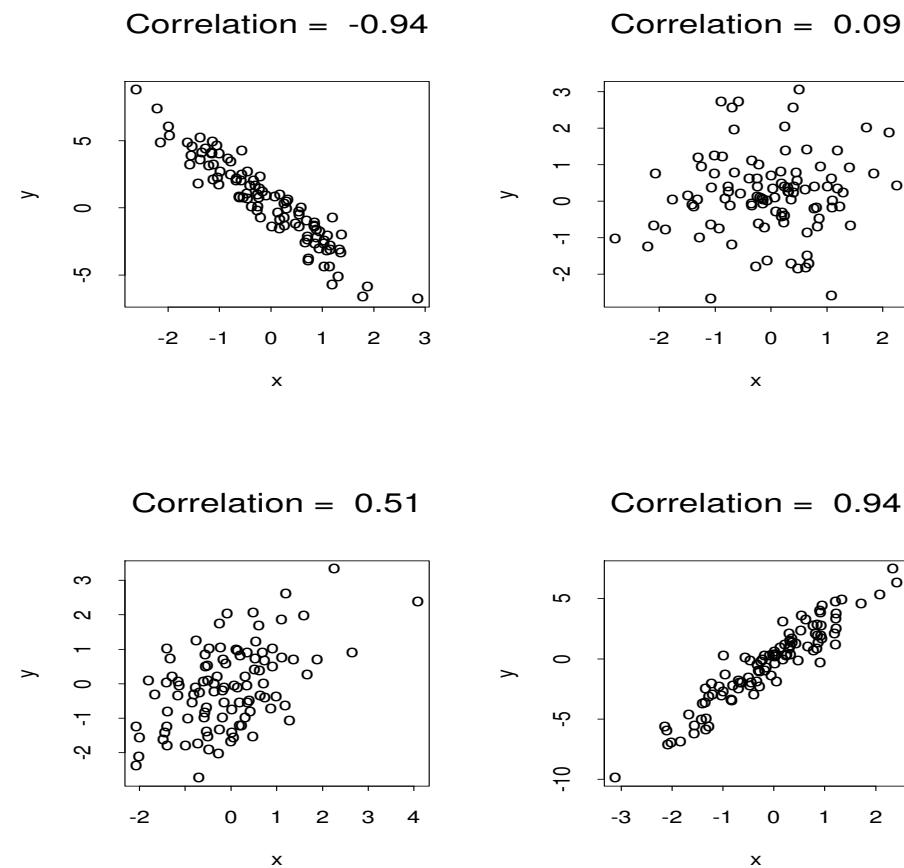


Figure 11.4: Examples of correlations

```
[1] -0.1563504
> cor(x1,x2)
[1] -0.1350433
> x3 = x1 + x2
> cov(x1,x2)
[1] -0.1563504
> cov(x1,x3)
[1] 1.042315
> cor(x1,x3)
[1] 0.6724712
>
```

The covariance and correlation matrices can also be obtained by the functions `cov()` and `cor()` using column vectors:

```
> cov(cbind(x1,x3))
      x1        x3
x1 1.198665 1.042315
x3 1.042315 2.004253
> cor(cbind(x1,x3))
      x1        x3
x1 1.0000000 0.6724712
x3 0.6724712 1.0000000
>
```

## 11.4 Scatter Plots in R

Scatter plots may be created with the `plot()` function. The simplest form is `plot(x,y)` where `x` and `y` are two numerical vectors of equal length, representing the horizontal and vertical axes respectively. It is also possible to use `plot(m)` where `m` is an  $n \times 2$  matrix, columns 1 and 2 representing the horizontal and vertical axes respectively.

We give a simple illustration. The following R code produces Figure 11.5.

```
> par(cex=1.0, cex.lab=1.25, cex.main = 1.5, mar=c(4,5,2,2))
> x = rnorm(100)
> y = x + rnorm(100)
> r = cor(x,y)
> ex0 = expression(paste('Title can include the Greek letter ',
+ beta,sep=''))
> ex1 = expression(italic(X)[1])
> ex2 = expression(italic(Y)[1])
> ex3 = bquote(italic(r) == .(round(r,2)) )
```

```
> plot(x,y,xlab=ex1,ylab=ex2, main=ex0)
> text(-2,2,ex3,cex=1.5)
```

A few features are worth noting:

1. The `par()` function is used for general plot settings. The `cex`, `cex.lab`, `cex.main` options are used to control the relative size of the plot elements. The `mar` option sets margin sizes. See `help(par)`.
2. A correlated sample is produced with the commands

```
> x = rnorm(100)
> y = x + rnorm(100)
> r = cor(x,y)
```

with the sample correlation stored in `r`.

3. The `expression()` and `bquote()` functions are used to create specialized plot text. Greek letters and subscripts are demonstrated. The `bquote()` function allows display of the contents of a variable.
4. The `text()` command places text at specific coordinates.
5. Useful functions also include `legend()`, `axis()`, `title()`, `box()` and `mttext()`.

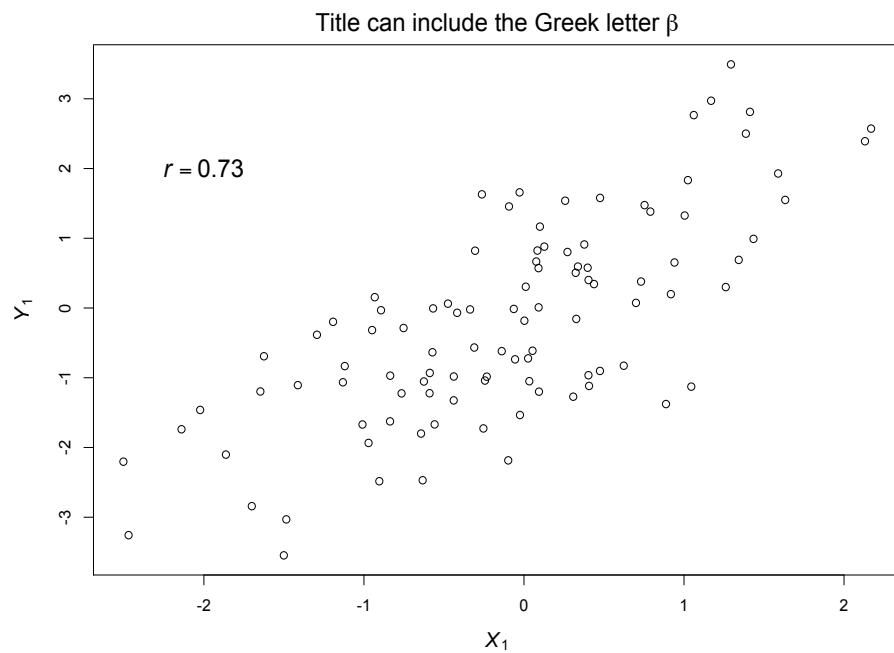


Figure 11.5: Creating scatter plot in R, Section 11.4.

# **Part III**

# **Statistics - Inference**

## Chapter 12

# Confidence Intervals and Hypothesis Tests - Population Mean

We are ready to discuss the subject of formal statistical inference. A typical opinion poll might that 37% of the population believe that water should not be fluoridated. The margin of error was  $+/- 3\%$  19 times out of 20. Note that there are 3 distinct elements to this statement.

1. The *estimate* is 37%.
2. The *margin of error* is 3%.
3. The *confidence level* is  $19/20 = 95\%$ .

The correct interpretation is that the *estimate* is within the *margin of error* of the true population value with a probability given by the *confidence level*. We now show how to construct such an inferential statement about a population mean based on a sample mean.

### 12.1 Confidence Intervals

We assume that there is a population of measurements and that we can, at least in principle, calculate a true population mean  $\mu$ . We further assume that doing so presents practical difficulties, so that we instead take a random sample from this population of size  $n$ , and accept the resulting sample mean  $\bar{X}_n$  as an estimate of  $\mu$ . Suppose that the true population variance is  $\sigma^2$ . Then we may define a  $(1 - \alpha)100\%$  confidence interval to be

$$CI_{1-\alpha} = \bar{X}_n \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

or, expressed in interval form

$$CI_{1-\alpha} = \left( \bar{X}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right).$$

Using the terminology we earlier defined

1. The *estimate* is  $\bar{X}_n$
2. The *margin of error* is  $z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ ,
3. The *confidence level* is  $1 - \alpha$ .

The inferential statement is that the true mean is within the margin of error of the estimate with a probability of  $1 - \alpha$ , the confidence level, or

$$P(\mu \in CI_{1-\alpha}) = 1 - \alpha$$

Usually, a confidence level is chosen in advance. Typical values are 90%, 95% or 99%, corresponding to values of  $\alpha$  of 0.10, 0.05, 0.01. The most commonly used is the 95% confidence level. In opinion polls, the expression *19 times out of 20* refers to a confidence level of 95%. This functions as a commonly accepted standard for evidence.

A selection of critical values is given below.

Confidence Level	$\alpha$	$z_{\alpha/2}$
90%	0.10	1.645
95%	0.05	1.960
99%	0.01	2.576

**Example 12.1** A process which produces ball bearings is known to have a standard deviation of 0.12 centimeters in terms of the diameter. A random sample of 50 is collected and a sample mean of 3.45 centimeters is calculated.

To construct a 95% confidence interval we first note that

$$\begin{aligned} n &= 50 \\ \bar{X}_{50} &= 3.45 \\ \alpha &= 0.05 \\ z_{\alpha/2} &= 1.96 \\ \sigma &= 0.12. \end{aligned}$$

Substituting these values into the formula for the confidence interval gives

$$\begin{aligned} CI_{95\%} &= \bar{X}_n \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \\ &= 3.45 \pm 1.96 \frac{0.12}{\sqrt{50}} \\ &= 3.45 \pm 0.03 \end{aligned}$$

giving a margin of error of 0.03. Alternatively, we can express the confidence interval in interval form:

$$\begin{aligned} CI_{95\%} &= (3.45 - 0.03, 3.45 + 0.03) \\ &= (3.42, 3.48). \end{aligned}$$
■

## 12.2 Hypothesis Tests

Another form of inference is the *hypothesis test*. Instead of trying to estimate a population mean  $\mu$  we test a statement about it.

**Example 12.2** A package of coffee is marked 300 grams. Suppose we purchase a package, and after measuring the contents we find that there are 298.8 grams. If we accept that our measuring error is normally distributed with mean 0 and standard deviation  $\sigma = 1$ , is the observed quantity consistent with the claim on the package?

The claim on the package is that  $\mu = 300$ , where  $\mu$  represents the true contents of the package. If that claim is true then if  $X$  is the contents of a randomly measured package,  $X \sim N(300, 1)$ . If we draw the distribution and note the location of the value 298.8, we see that this observed value is quite consistent with the claim. On the other hand, if the measured quantity was 296, we would suspect that the claim was not true. ■

When we study hypothesis testing we are formalizing the concepts illustrated in the previous example. A hypothesis test consists of three distinct parts.

1. We begin with a *null hypothesis*  $H_o$  and an *alternative hypothesis*  $H_a$ . The null hypothesis represents the *status quo*, or the state of things we accept as true if there is no evidence to the contrary.
2. A *test statistic*. This is the summary of the available data in a form convenient for testing.
3. The *level of significance* is the degree to which the evidence supports  $H_a$ . The smaller the level of significance, the greater the evidence in favor of  $H_a$ .

If the hypothesis test concerns a population mean, we begin with some fixed hypothetical value  $\mu_0$ . In the previous example we had  $\mu_0 = 300$ . Then there are three kinds of hypothesis sets

1. One sided, lower tailed hypothesis

$$\begin{aligned} H_o : \mu &\geq \mu_0 \\ H_a : \mu &< \mu_0 \end{aligned}$$

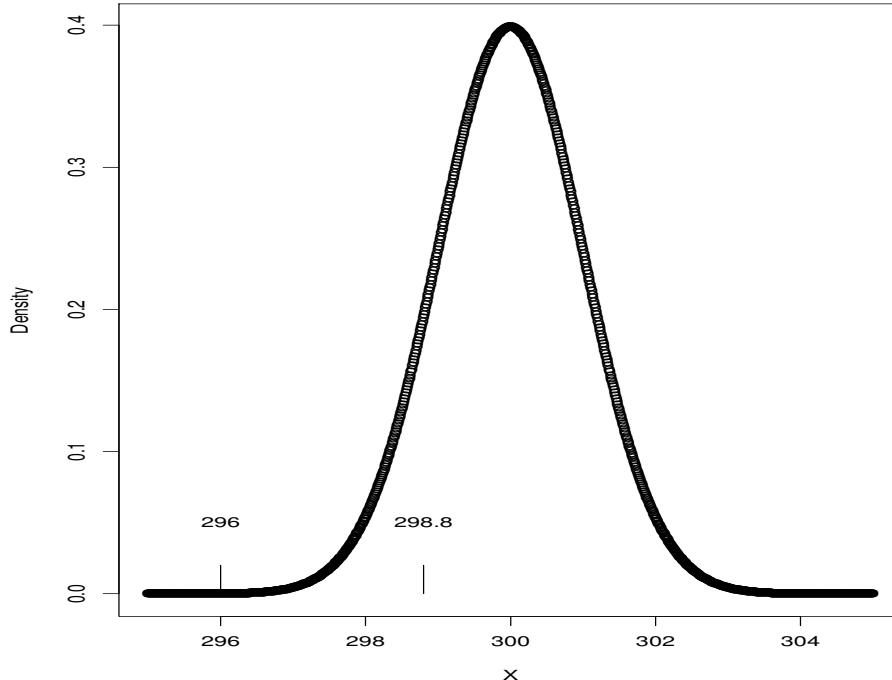


Figure 12.1: Location of observed quantities relative to the normal density

We are looking for evidence that the true mean is *less than* the hypothetical mean.

## 2. One sided, upper tailed hypothesis

$$H_o : \mu \leq \mu_0$$

$$H_a : \mu > \mu_0$$

We are looking for evidence that the true mean is *greater than* the hypothetical mean.

## 3. Two sided hypothesis

$$H_o : \mu = \mu_0$$

$$H_a : \mu \neq \mu_0$$

We are looking for evidence that the true mean is *not equal to* the hypothetical mean.

Given a random sample  $X_1, \dots, X_n$  from the population the *test statistic* we use is

$$Z_{obs} = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}.$$

The calculation of  $\alpha_{obs}$ , the *observed level of significance*, depends on the kind of hypothesis test. If we let  $Z$  be any standard normal random variable, the we use the following formulae:

1. One sided, lower tailed hypothesis

$$\alpha_{obs} = P(Z < Z_{obs})$$

2. One sided, upper tailed hypothesis

$$\alpha_{obs} = P(Z > Z_{obs})$$

3. Two sided hypothesis

$$\begin{aligned}\alpha_{obs} &= P(Z < -|Z_{obs}| \text{ OR } |Z_{obs}| < Z) \\ &= 2P(Z > |Z_{obs}|)\end{aligned}$$

If we have  $0.10 \geq \alpha_{obs} > 0.05$  we generally say there is weak evidence to reject  $H_o$  in favor of  $H_a$ . If  $\alpha_{obs} \leq 0.05$  we say there is strong evidence to reject  $H_o$  in favor of  $H_a$ . Sometimes, a significance level  $\alpha$  is determined in advance. If we calculate  $\alpha_{obs} \leq \alpha$ , we conclude that the null hypothesis can be rejected at a significance level of  $\alpha$ . A common choice for  $\alpha$  is 0.05. The observed level of significance is commonly known as the *P-value*.

**Example 12.3** To continue the previous example, suppose  $n = 10$  packages of coffee are sampled, and that the resulting mean is 298.4 grams. Is there evidence that the packages are being underfilled?

We are looking specifically for evidence that the packages are underfilled, so we have a one sided lower tailed hypothesis. Recall that  $\mu_0 = 300$ ,  $\sigma = 1$ . The hypotheses are then

$$\begin{aligned}H_o &: \mu \geq 300 \\ H_a &: \mu < 300.\end{aligned}$$

The test statistic is

$$\begin{aligned}Z_{obs} &= \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \\ &= \frac{298.4 - 300}{1/\sqrt{10}} \\ &= -5.06.\end{aligned}$$

Using the appropriate formula we get an observed significance level of

$$\begin{aligned}\alpha_{obs} &= P(Z < Z_{obs}) \\ &= P(Z < -5.06) \\ &< 0.0002\end{aligned}$$

Note that we cannot directly calculate the probability from the tables since  $-5.06 < -3.49$ , the lower limit. We can say the the observed level of significance is less than 0.0002, the smallest value represented on the table. This gives very strong evidence against  $H_o$ . ■

**Example 12.4** To continue the example, suppose the flaw in the claim of 300 grams is brought to the attention of the manufacturer. They would then recalibrate their machines to ensure the correct quantity. A new sample would be taken. Suppose the sample is of size  $n = 20$  and the observed sample mean is 300.1 grams. So a new hypothesis test is done, except that the manufacturer is looking for evidence of either underfilling or overfilling, that is, they are looking for evidence that either  $\mu > 300$  or  $\mu < 300$ . In this case a two-sided test is appropriate, giving hypotheses

$$\begin{aligned}H_o : \mu &= 300 \\ H_a : \mu &\neq 300.\end{aligned}$$

The test statistic is then

$$\begin{aligned}Z_{obs} &= \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \\ &= \frac{300.1 - 300}{1/\sqrt{20}} \\ &= 0.45\end{aligned}$$

and the observed significance level is

$$\begin{aligned}\alpha_{obs} &= 2P(Z > |Z_{obs}|) \\ &= 2P(Z > 0.45) \\ &= 2(1 - 0.6736) \\ &= 0.6528.\end{aligned}$$

Since  $\alpha_{obs} > 0.1$  we can conclude that there is no evidence against the null hypothesis that  $\mu = 300$  grams. ■

### 12.3 The *t*-distribution

The procedures for calculating confidence intervals and performing hypothesis test previously discussed share one critical flaw. Namely, it is rare that

we actually know the value of the variance  $\sigma^2$ , yet the calculations require that we know it.

Of course, we could simply substitute the sample variance  $S^2$ . The problem with this is that we introduce additional sampling error into our calculations, and so we run the risk of generally overestimating the accuracy of our inference. Fortunately, it is quite simple to adjust for this. This adjustment is based on the fact that the distribution of

$$T = \frac{\bar{X}_n - \mu}{S_n / \sqrt{n}}$$

has a  $t$ -distribution with  $n - 1$  degrees of freedom, assuming that the sample  $X_1, \dots, X_n$  from which  $\bar{X}_n$  and  $S_n^2$  are calculated is a random sample from a normal distribution with mean  $\mu$  (see Section 4.15).

We explain why this is the case. Suppose  $X_1, \dots, X_n$  is an *iid* sample from normal distribution  $N(\mu, \sigma^2)$ . The sample variance is written:

$$S_n^2 = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{n - 1}.$$

If we set

$$W = \frac{(n - 1)S_n^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\sigma^2},$$

then it may be shown that

$$W \sim \chi_{n-1}^2,$$

that is,  $W$  has a  $\chi^2$  distribution with  $n - 1$  degrees of freedom (see Section 4.15). Next, we note that

$$\bar{X}_n \sim N(\mu, \sigma^2/n),$$

so that

$$Z = \frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}} \sim N(0, 1).$$

It may also be shown that  $Z \perp W$ . This means

$$\begin{aligned} T &= \frac{\bar{X}_n - \mu}{S_n / \sqrt{n}} \times \frac{1/\sigma}{1/\sigma} \\ &= \frac{\frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}}}{\sqrt{\frac{(n-1)S_n^2}{(n-1)\sigma^2}}} \\ &= \frac{Z}{\sqrt{W/(n-1)}}, \end{aligned}$$

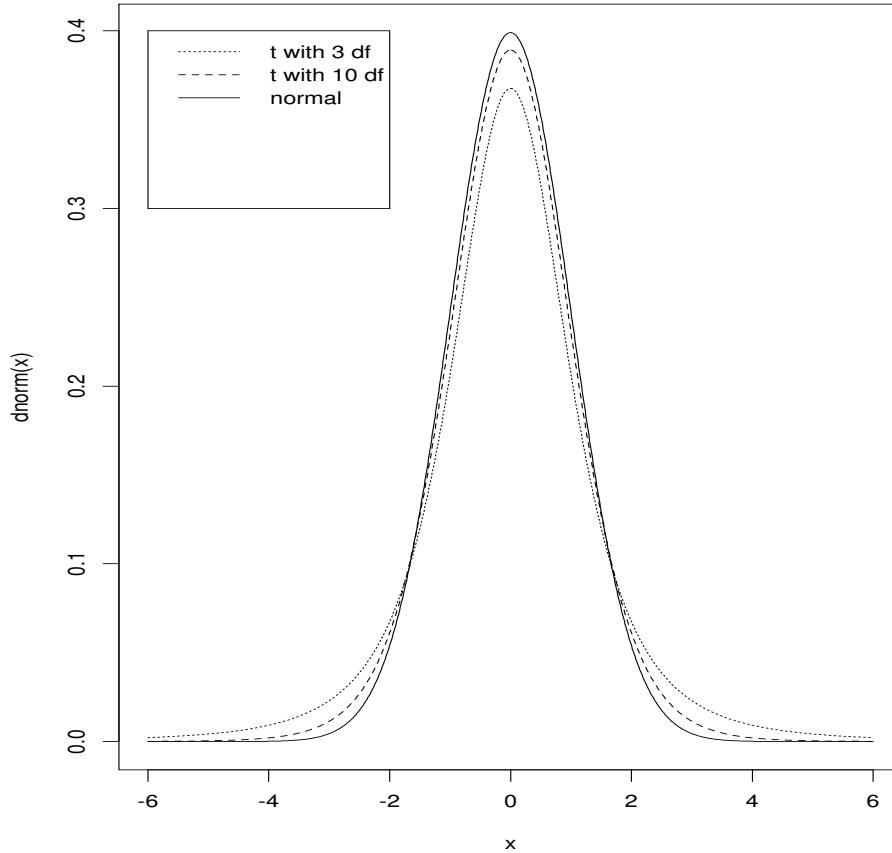


Figure 12.2: Several  $t$ -distribution density functions

which matches the definition of the  $t$ -distribution with  $n - 1$  degrees of freedom from Section 4.15. The  $t$ -distribution is widely used, and we can obtain  $\alpha$  critical values  $t_{n-1,\alpha}$  from Table A.3.

As we might expect, the  $t$ -distribution resembles a normal distribution, except that there is somewhat more variability. As the degrees of freedom increase, the resemblance to the normal becomes greater. Note that the last line of the table is labeled as  $df = \infty$ . This corresponds to the standard normal distribution.

### 12.3.1 Calculating Confidence Intervals with Unknown Variance

If we have a random sample  $X_1, \dots, X_n$  from a population with mean  $\mu$  and variance  $\sigma^2$  we can calculate a level  $1 - \alpha$  confidence interval using the formula

$$CI_{1-\alpha} = \bar{X}_n \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

If  $\sigma^2$  is unknown, then we substitute  $S_n$  for  $\sigma$  and  $t_{n-1,\alpha/2}$  for  $z_{\alpha/2}$  giving

$$CI_{1-\alpha} = \bar{X}_n \pm t_{n-1,\alpha/2} \frac{S_n}{\sqrt{n}}.$$

Note that the principal effect of the substitution is to make the confidence interval wider, in order to reflect the greater uncertainty inherent in using an estimate for  $\sigma^2$  instead of the true value.

**Example 12.5** If we collect a sample of size  $n = 5$  from a population with mean  $\mu$ , and obtain

$$\begin{aligned}\bar{X}_5 &= 34.6 \\ S_5 &= 5.7\end{aligned}$$

then to calculate a 95% confidence interval we read the critical value

$$\begin{aligned}t_{n-1,\alpha/2} &= t_{4,0.025} \\ &= 2.7764\end{aligned}$$

from Table A.3. The confidence interval is then

$$\begin{aligned}CI_{.95} &= \bar{X}_n \pm t_{n-1,\alpha/2} \frac{S_n}{\sqrt{n}} \\ &= 34.6 \pm 2.7764 \frac{5.7}{\sqrt{5}} \\ &= 34.6 \pm 7.08.\end{aligned}$$
■

### 12.3.2 Hypothesis Tests with Unknown Variance

A similar procedure can be used for hypothesis tests involving population means when the variance is unknown. The structure of the hypotheses remains unchanged.

The test statistics becomes

$$T_{obs} = \frac{\bar{X}_n - \mu_0}{S_n / \sqrt{n}}$$

where  $\mu_0$  is the hypothetical mean. To calculate the observed significance level we use the formulae

1. One sided, lower tailed hypothesis

$$\alpha_{obs} = P(T_{n-1} < T_{obs})$$

2. One sided, upper tailed hypothesis

$$\alpha_{obs} = P(T_{n-1} > T_{obs})$$

3. Two sided hypothesis

$$\begin{aligned}\alpha_{obs} &= P(T_{n-1} < -|T_{obs}| \text{ OR } |T_{obs}| < T_{n-1}) \\ &= 2P(T_{n-1} > |T_{obs}|)\end{aligned}$$

where  $T_{n-1}$  is a random variable with a  $t$ -distribution with  $n - 1$  degrees of freedom.

**Example 12.6** Suppose we observe a sample of 5 earthworms with weights in ounces

1.4    1.5    1.7    1.9    2.1.

It is suspected that the average earthworm weighs more than last years mean of 1.55 ounces. Is there evidence to support that hypothesis?

The hypothesis is expressed as a one sided upper tailed hypothesis, giving hypotheses

$$\begin{aligned}H_o : \mu &\leq 1.55 \\ H_a : \mu &> 1.55\end{aligned}$$

where  $\mu$  is the true population mean weight, and  $\mu_0 = 1.55$  is the hypothetical mean weight.

We have

$$\begin{aligned}\bar{X}_5 &= 1.72 \\ S_5 &= 0.286\end{aligned}$$

giving the test statistic

$$\begin{aligned}T_{obs} &= \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} \\ &= \frac{1.72 - 1.55}{0.286/\sqrt{5}} \\ &= 1.329\end{aligned}$$

To calculate the observed significance level, we note that there are  $n - 1 = 4$  degrees of freedom. From Table A.3 we note that

$$\begin{aligned} t_{4,0.25} &= 0.741 \\ t_{4,0.10} &= 1.533 \end{aligned}$$

The observed significance level is calculated by

$$\begin{aligned} \alpha_{obs} &= P(T_{n-1} > T_{obs}) \\ &= P(T_{n-1} > 1.329) \end{aligned}$$

From the critical values obtained from the table, we can conclude that

$$\begin{aligned} P(T_{n-1} > 0.741) &= 0.25 \\ P(T_{n-1} > 1.533) &= 0.10. \end{aligned}$$

Then since

$$0.741 < T_{obs} < 1.533$$

we must have

$$0.10 < \alpha_{obs} < 0.25.$$

We therefore conclude that we have no evidence that the population mean weight is larger than 1.55 ounces. ■

**Example 12.7** If in the previous example we were to use a two-sided hypothesis

$$\begin{aligned} H_o : \mu &= 1.55 \\ H_a : \mu &\neq 1.55 \end{aligned}$$

then since the observed significance level for a two sided test is

$$\begin{aligned} \alpha_{obs} &= P(T_{n-1} < -|T_{obs}| \text{ OR } |T_{obs}| < T_{n-1}) \\ &= 2P(T_{n-1} > |T_{obs}|) \end{aligned}$$

we would obtain

$$2 \times 0.10 < \alpha_{obs} < 2 \times 0.25. \quad \blacksquare$$

## 12.4 Assumptions

The techniques described in this section all depend to some degree on some assumptions.

First of all, the sample is assumed to be a random sample. This means that every possible sample has an equal chance of being selected. Another

way of putting this is that the probability that a given unit from the population is sampled is unaffected by whether or not any other particular unit is sampled. For example, the probability that I am selected is not affected by whether or not my brother was selected. Care is sometimes needed to ensure that this condition is met.

Second of all, the probability calculations depended on some distributional assumptions. When the population variance  $\sigma^2$  is known, the central limit theorem is usually sufficient to justify using the standard normal distribution. A rule of thumb is that when  $n < 30$  some caution with respect to this assumption is in order.

When the variance is unknown, the assumption behind the use of the  $t$ -distribution is that the population distribution itself is normally distributed. In practice, inference using the  $t$ -distribution is reasonably accurate when the data is not too skewed.

## Chapter 13

# Some General Definitions for Hypothesis Testing

We have seen how a hypothesis test works for the population mean in Chapter 12. The idea, however, is quite a general one, so it is necessary to understand the underlying principles. The first idea is that of the *parameter*:

**Definition 13.1** A *parametric family of distributions* is a class of distributions  $P_\theta$  indexed by one or more parameters, denoted  $\theta$ . The parameter is assumed to belong to a *parameter space*  $\Theta$ .

The parameter is the object of investigation. Any quantity used to specify a distribution which is known is *fixed*. Any such quantity which is unknown, but not the subject of inquiry is a *nuisance parameter*.

Any precise statement about an unknown parameter is known as an *inference*. An inference must include the probability that the statement is wrong. ■

We have seen in Chapter 4 a number of distributions which require one or more values to define. We next review some of the more common examples.

**Example 13.1** A binomial distribution  $bin(n, p)$  requires two numbers to be completely specified. The number of respondents  $X$  who support a proposition in an opinion poll is often modeled as a binomial random variable. However, the quantity  $n$ , in this case the sample size, is usually known, or *fixed*. Then,  $\theta = p$  is the parameter. We may set the parameter space to be  $\Theta = [0, 1]$ . However, if we may restrict the possible values of  $p$  to a strict subset of  $[0, 1]$ , this becomes the parameter space. ■

**Example 13.2** A normal distribution  $N(\mu, \sigma^2)$  is defined by two quantities  $\mu, \sigma^2$ . Usually, though not always, interest is in the mean  $\mu$ . In this case the analysis is simplified if  $\sigma^2$  is fixed, so the parameter is  $\theta = \mu$ . The parameter

space can be  $\Theta = (-\infty, \infty)$ , but is often smaller. For example, we may rule out negative values of  $\mu$ , in this case  $\Theta = [0, \infty)$ .

Of course, it is usually unrealistic to assume  $\sigma^2$  is known, and so is usually a nuisance parameter. In this case the parameter  $\theta = (\mu, \sigma^2)$  is two dimensional, and the parameter space may be set as  $\Theta = (-\infty, \infty) \times [0, \infty)$ . ■

There are two main forms of inference. If the object is to estimate a parameter  $\theta$  we rely on a *confidence interval*.

**Definition 13.2** Suppose we are given a parametric family of distributions, with parameter  $\theta \in \Theta$ . Let  $\tilde{X}$  be a sample from one member of the parametric family, specified by *true value*  $\theta_0$ .

Let  $\theta_i$  be the  $i$ th component of  $\theta$ . A *confidence interval* of *confidence level*  $1 - \alpha$  for  $\theta_i$  is any interval

$$CI_{1-\alpha} = (L, U)$$

constructed from the data  $\tilde{X}$  which has the property

$$P(\theta_i \in CI_{1-\alpha}) = P(L \leq \theta_i \leq U) \leq \alpha. \quad (13.1)$$

The confidence cannot depend on any component of  $\theta$ . If inequality in (13.1) can be replaced by equality, the confidence interval is exact. If inequality cannot be ruled out, the confidence interval is *conservative*. The actual probability  $P(\theta_i \in CI_{1-\alpha})$  is referred to as the *coverage*. When we may claim (13.1), then  $1 - \alpha$  is known as the *nominal coverage* or confidence level. ■

It is often the case, especially with more challenging inference problems, that confidence intervals are conservative. In such cases, there is usually some investigation in methodological journals into the difference between the actual and nominal coverage.

The object of estimation based on confidence intervals is to identify, with as great a resolution as possible, the distribution from which the data  $\tilde{X}$  was sampled. Although it might seem that this is sufficient to resolve any inference problem, it is best to recognize that inference usually deals with one of two general questions:

1. From what distribution was  $\tilde{X}$  sampled?
2. Could  $\tilde{X}$  have been sampled from *this* distribution?

We have already seen, informally, a number of problems of the second type. In Example 8.4 we asked whether or not it was unusual that 4 of 5 children in a family are male. Technically, we are asking if  $X = 4$  is

compatible with a  $\text{bin}(5, 1/2)$  distribution. In the maze example of Section 7.2.1 we developed a model of maze navigation based on the “memoryless” hypothesis. We could then compare the distributional properties of this model to experimental observations. If the observations are not compatible with the model, we would reject the “memoryless hypothesis”.

The technical problem underlying the theory of hypothesis testing is to define precisely what is meant by “compatible with the model”. Suppose we conduct a survey in which respondents are asked which of two presidential candidates they support. Actual elections results rarely diverge greatly from a 50% split (a victory of 55% to 45% is considered decisive). If we sample  $n = 2000$  respondents, it is possible, at least in theory, that all 2000 respondents will support the same candidate, since that candidate will have at least 2000 supporters in the population being sampled. It is not hard to determine that this probability is negligible ( $= 2 \times (1/2)^{2000}$ ), but we are left with the problem of deciding which observations are or are not “compatible” with a probability model. This makes clear the probabilistic nature of inference, and the need to report an error probability.

In contrast to estimation, a hypothesis test is intended to answer a specific question regarding an unknown parameter. For example, we may wish to know whether or not the level of support for a candidate is above 50%. The question takes the form of competing hypotheses:

**Definition 13.3** Suppose we are given a parametric family of distributions as defined in 13.1, and data  $\tilde{X}$  from true distribution  $\theta_0 \in \Theta$ . Suppose we are given a strict subset of the parameter space  $\Theta_o \subset \Theta$ , and let  $\text{Theta}_a = \Theta - \Theta_o$ , so that  $\Theta_o$  and  $\Theta_a$  partition  $\Theta$ . This permits us to define a *null hypothesis* and an *alternative hypothesis*:

$$H_o : \theta \in \Theta_o \text{ and } H_a : \theta \in \Theta_a.$$

The object of a hypothesis test is to determine whether or not the data  $\tilde{X}$  is compatible with some distribution  $\theta \in \Theta_o$ . If not, we *reject* the null hypothesis  $H_o$ , otherwise we *accept* the null hypothesis. A hypothesis consisting of a single parameter is a *simple hypothesis*, and one consisting of multiple parameters is a *composite hypothesis*. ■

The careful reader will have noted the guarded language of Definition 13.3. The object is not to determine whether the true parameter  $\theta_0$  is in  $\Theta_o$  or  $\Theta_a$  (that problem is more accurately referred to as *classification*). The structure of the problem is not symmetric with respect to the hypotheses. The null hypothesis plays a special role, representing the state of affairs we accept in the absence of evidence. For example, in a criminal trial, if absolutely no evidence is presented, the decision would be *acquittal*, meaning *not guilty*. Hence, the null hypothesis is that the defendant did not commit

the crime. This is why a decision to acquit is appropriately denoted *not guilty* rather than *innocent*, the burden of proof resting with the prosecution.

Similarly, in a hypothesis test, we proceed by attempting to prove that the null hypothesis  $H_o$  is not correct. This is done statistically by investigating whether or not the data  $\tilde{X}$  is compatible with any distribution  $P_\theta$  for which  $\theta \in \Theta_o$ . If it is, we cannot reject  $H_o$ , otherwise, we reject  $H_o$  in favor of  $H_a$ . We could, in principle, consider the degree to which the data constitutes evidence of  $H_o$ , but this would constitute a new problem.

### 13.1 Hypothesis Testing Based on Rejection Regions

In Section 12.2 the hypothesis test was based on the intuitive observation that questions concerning the parameter  $\mu$  could be resolved by observing the sample mean  $\bar{X}_n$ , being close to  $\mu$  within a predictable error tolerance. If we wish to characterize evidence against null hypothesis  $H_o : \mu \leq \mu_0$ , then we would take large values of  $\bar{X}_n$  to constitute such evidence.

However, we can distinguish between two approaches. In Section 12.2, no formal decision to accept or reject the null hypothesis  $H_o$  was proposed. Rather, a degree of evidence with which to reject  $H_o$  was calculated. The *rejection region* defines a formal decision rule.

**Example 13.3** Suppose we are given a parametric family of distributions, with parameter  $\theta \in \Theta$ . Let  $\tilde{X}$  be a sample from one member of the parametric family, specified by *true value*  $\theta_0 \in \Theta$ . Let  $H_o : \theta \in \Theta_o$  and  $H_a : \theta \in \Theta_a$  be null and alternative hypotheses.

Let  $\mathcal{X}$  to be the set of all possible samples that could be collected. Suppose  $\mathcal{X}$  is partitioned into two groups, *rejection region*  $R$  and acceptance region  $A = \mathcal{X} - R$ . The intention is that  $H_o$  is rejected if  $\tilde{X} \in R$ . If

$$P_\theta(\tilde{X} \in R) \leq \alpha \text{ for all } \theta \in \Theta_o$$

then  $\alpha$  is a *level of significance* for the test or rejection region. If

$$\sup_{\theta \in \Theta_o} P_\theta(\tilde{X} \in R) = \alpha$$

then  $\alpha$  is the *size* of the test or rejection region. ■

#### 13.1.1 Rejection Regions for Testing a Single Population Mean with Known Variance

We will apply the concept of the rejection region to the upper tailed hypothesis test for a population mean when the variance is known. The hypothesis

is

$$\begin{aligned} H_o &: \mu \leq \mu_0 \\ H_a &: \mu > \mu_0 \end{aligned}$$

where  $\mu_0$  is some hypothetical mean. The observed level of significance is

$$\alpha_{obs} = P(Z > Z_{obs})$$

where

$$Z_{obs} = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}.$$

If we specify a significance level  $\alpha$  in advance, then we reject  $H_o$  if

$$\alpha_{obs} \leq \alpha.$$

By definition of  $\alpha_{obs}$  this condition is equivalent to

$$Z_{obs} \geq z_\alpha.$$

This then defines the rejection region

$$R = \{ \text{All samples for which } Z_{obs} \geq z_\alpha \},$$

meaning that we reject  $H_o$  at an  $\alpha$  significance level if  $Z_{obs} \geq z_\alpha$ . This should make sense intuitively. The larger the value of  $\bar{X}_n$  the more we are likely to conclude that  $H_o$  is not true. Since  $Z_{obs}$  increases with  $\bar{X}_n$ , we can see that the hypothesis test is equivalent to rejecting  $H_o$  if  $\bar{X}_n$  is above a certain value. In fact the rejection region can be rewritten

$$\bar{X}_n \geq \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}.$$

There will be three different forms of the rejection region depending on the form of the hypothesis.

1. One sided, lower tailed hypothesis

Reject  $H_o$  if  $Z_{obs} \leq -z_\alpha$

2. One sided, upper tailed hypothesis

Reject  $H_o$  if  $Z_{obs} \geq z_\alpha$

3. Two sided hypothesis

Reject  $H_o$  if  $|Z_{obs}| \geq z_{\alpha/2}$

**Example 13.4** We will redo Example 9.4 using the rejection region. Recall that we had hypothesis

$$\begin{aligned}H_o &: \mu = 300 \\H_a &: \mu \neq 300.\end{aligned}$$

with test statistic

$$\begin{aligned}Z_{obs} &= \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \\&= \frac{300.1 - 300}{1/\sqrt{20}} \\&= 0.45.\end{aligned}$$

If we test using a significance level  $\alpha = 0.05$  then the rejection region is

$$\begin{aligned}|Z_{obs}| &\geq z_{\alpha/2} \\&= z_{0.025} \\&= 1.96.\end{aligned}$$

We have  $|Z_{obs}| = 0.45$ , so that  $|Z_{obs}| < z_{0.025}$ . We therefore conclude that there is no evidence to reject  $H_o$  at a significance level of  $\alpha = 0.05$ . ■

### 13.1.2 Rejection Regions for Testing a Single Population Mean with Unknown Variance

The role of the rejection region is exactly the same when testing for a population mean with unknown variance, except that the rejection regions are constructed using critical values from the  $t$ -distribution with  $n - 1$  degrees of freedom, where  $n$  is the sample size used.

1. One sided, lower tailed hypothesis

Reject  $H_o$  if  $T_{obs} \leq -t_{n-1,\alpha}$

2. One sided, upper tailed hypothesis

Reject  $H_o$  if  $T_{obs} \geq t_{n-1,\alpha}$

3. Two sided hypothesis

Reject  $H_o$  if  $|T_{obs}| \geq t_{n-1,\alpha/2}$

## 13.2 Type I and Type II Errors

When rejection regions are used for hypothesis testing it becomes useful to introduce two types of errors that can occur.

**Definition 13.4** For a hypothesis test a *type I error* occurs when  $H_o$  is rejected when  $H_o$  is actually true. A *type II error* occurs when  $H_o$  is not rejected when  $H_a$  is actually true. ■

Consider again the upper tailed test for a population mean when  $\sigma^2$  is known. Given significance level  $\alpha$ , we reject  $H_o$  if

$$Z_{obs} \geq z_\alpha.$$

If  $H_o$  is true, and the true mean equals the hypothetical mean  $\mu_0$ , then  $Z_{obs}$  has a standard normal distribution, so that

$$\begin{aligned} P(\text{ Type I error }) &= P(Z_{obs} \geq z_\alpha | H_o) \\ &= \alpha. \end{aligned}$$

Note that if  $H_o$  is composite the probability of a type I error will depend on the particular value of  $\theta \in \Theta_o$ . The convention is to use the value which maximizes this probability, which will almost always be the value on the boundary between the null and alterative hypotheses.

This means that if the test uses a level  $\alpha$  rejection region, then the probability of a type I error is  $\alpha$ . This will be true for any properly constructed hypothesis test based on rejection regions. In this context, the predetermined significance level is often referred to as the *size* of the test. This means that  $\alpha$  can be interpreted as the largest probability of falsely rejecting  $H_o$  if  $H_o$  is true that we would be willing to tolerate.

## 13.3 Power

Again consider the upper tailed alternative hypothesis

$$\begin{aligned} H_o : \mu &\leq \mu_0 \\ H_a : \mu &> \mu_0 \end{aligned}$$

A Type II error occurs when  $H_o$  is not rejected when  $H_a$  is true. If we wish to calculate the probability of a Type II error, we must specify exactly which value of  $\mu > \mu_0$  is the correct one. Usually, the notation

$$\begin{aligned} \beta(\mu) &= P(\text{ Type II error }) \\ &= P(Z_{obs} < z_\alpha | \mu) \end{aligned}$$

is used. We then define

$$\begin{aligned}\text{Power}(\mu) &= 1 - \beta(\mu) \\ &= P(\text{correctly rejecting } H_o \text{ when the true mean is } \mu)\end{aligned}$$

again noting that the exact value of the power depends on the exact value of  $\mu$ .

A hypothesis test is usually designed to have a fixed size  $\alpha = 0.05$ , although other smaller or larger values of  $\alpha$  may be appropriate, depending on the application. Given a fixed size, the only way to increase the power is to increase the sample size. This means that an appropriate sample size is usually determined by calculating power.

**Example 13.5** Consider testing

$$\begin{aligned}H_o : \mu &\leq 100 \\ H_a : \mu &> 100\end{aligned}$$

when the population standard deviation is  $\sigma = 10$ . Suppose a sample of size  $n = 50$  is used. What is the power of a size  $\alpha = 0.05$  test when the actual mean is  $\mu = 105$ ?

First note that when  $\mu = 105$  we have

$$\bar{X} \sim N\left(105, \frac{10^2}{50}\right).$$

The test statistic is

$$\begin{aligned}Z_{obs} &= \frac{\bar{X} - 100}{\sqrt{10^2/50}} \\ &= \frac{\bar{X} - 100}{1.414}\end{aligned}$$

A size  $\alpha = 0.05$  rejects  $H_o$  if

$$Z_{obs} > z_{0.05}$$

or

$$\bar{X} > 100 + z_{0.05} \times 1.414.$$

This leads to

$$\begin{aligned}
 \text{Power}(105) &= P(\bar{X} > 100 + z_{0.05} \times 1.414) \\
 &= P\left(\frac{\bar{X} - 105}{1.414} > \frac{100 + z_{0.05} \times 1.414 - 105}{1.414}\right) \\
 &= P\left(Z > \frac{-5 + z_{0.05} \times 1.414}{1.414}\right) \\
 &= P(Z > -3.54 + z_{0.05}) \\
 &= P(Z > -3.54 + 1.645) \\
 &= P(Z > -1.89) \\
 &\approx 0.9706
 \end{aligned}$$

■

### 13.4 Precise Definition of the Observed Level of Significance

The observed level of significance has been treated as type of index giving the degree to which the evidence contradicts the null hypothesis  $H_o$  (with smaller values tending to disprove  $H_o$ ). There exists a more precise characterization:

**Definition 13.5** The *observed level of significance* or *P-value* is the probability that a new sample collected under identical conditions would be more contradictory of the null hypothesis  $H_o$  than the one observed, assuming  $H_o$  is true. ■

This means that if the observed level of significance is very small, then it is unlikely that this sample would be seen if  $H_o$  is true.

## Chapter 14

# Inference for Differences of Means

In the previous section, we concerned ourselves with the problem of how to conduct inference for a single population mean  $\mu$ . It is frequently the case, however, that we are interested in the difference between two means  $\mu_1$  and  $\mu_2$ . The object is then to say something about  $\mu_2 - \mu_1$ , rather than to say something about the means taken independently.

There are two kinds of samples that can be used to draw inference about  $\mu_2 - \mu_1$ .

1. *Independent samples.* We have two statistically independent samples. The first is drawn from a population with mean  $\mu_1$ , and the second from a population with mean  $\mu_2$ .
2. *Paired samples.* In this case, a single sample of units is drawn from a single population. Two kinds of measurements are made on each unit. The mean of the first kind of measurement in the population is  $\mu_1$ , and the mean of the second kind of measurement in the population is  $\mu_2$ .

**Example 14.1** Suppose we are interested in the question of whether or not there is a difference in the amount of time spent watching television between men and women. Suppose 500 men and 500 women are sampled, and the two samples are collected independently of each other. If  $\mu_1$  and  $\mu_2$  are the average number of hours spent per day watching television for men and women in the population respectively, then we wish to know whether or not  $\mu_2 - \mu_1 = 0$ . In this case we have two independent samples. ■

**Example 14.2** If we are interested in whether or not a cholesterol reduction program is effective, we would select a sample of subjects from a well defined population, measure the subject's cholesterol level, apply the program, then

measure again the subject's cholesterol. If we imagine that we could apply the program to all members of the defined population, then  $\mu_1$  would be the population mean of the pre-program cholesterol levels, and  $\mu_2$  would be the population mean of the post-program cholesterol levels. This is an example of a paired sample. This difference here is that each subject provides both measurements. ■

**Example 14.3** To return to the television watching time example, we might find it less costly to sample 500 households and ask both the husband and wife for the amount of time spent watching television. (assuming, of course, that it is known that there are both in the household). In this case, we could not treat the sample of men and women as independent. It would therefore be a paired sample. ■

We now look at each kind of sample separately.

## 14.1 Independent Samples

We assume that we have two populations, and that a random sample is drawn from each. We assume that the mechanisms for drawing the samples are statistically unrelated. We then have the following quantities associated with this process.

	Pop'n 1	Pop'n 2
Population mean	$\mu_1$	$\mu_2$
Population variance	$\sigma_1^2$	$\sigma_2^2$
Sample size	$n_1$	$n_2$
Sample mean	$\bar{X}_1$	$\bar{X}_2$
Sample variance	$S_1^2$	$S_2^2$

As with the single population mean we can define three type of hypotheses, based on a hypothetical difference in mean

$$\mu_2 - \mu_1 = \Delta.$$

In most applications we have  $\Delta = 0$ , but the test works essentially the same way for any value.

1. One sided, lower tailed hypothesis

$$H_o : \mu_2 - \mu_1 \geq \Delta$$

$$H_a : \mu_2 - \mu_1 < \Delta$$

We are looking for evidence that  $\mu_2 - \mu_1$  is *less than*  $\Delta$ .

2. One sided, upper tailed hypothesis

$$\begin{aligned} H_o &: \mu_2 - \mu_1 \leq \Delta \\ H_a &: \mu_2 - \mu_1 > \Delta \end{aligned}$$

We are looking for evidence that  $\mu_2 - \mu_1$  is *greater than*  $\Delta$ .

3. Two sided hypothesis

$$\begin{aligned} H_o &: \mu_2 - \mu_1 = \Delta \\ H_a &: \mu_2 - \mu_1 \neq \Delta \end{aligned}$$

We are looking for evidence that  $\mu_2 - \mu_1$  is *not equal to*  $\Delta$ .

The exact technique we use will depend on what assumptions we can make about the variances. We consider three cases.

**Case 1** Both population variance  $\sigma_1^2$  and  $\sigma_2^2$  are known.

**Case 2** The population variances are unknown, but we assume they are equal, so that  $\sigma_1^2 = \sigma_2^2$ .

**Case 3** The population variances are unknown and we cannot assume that they are equal.

We now consider the three cases separately.

#### 14.1.1 Case 1 - Variances are Known

In this case, the  $(1 - \alpha)100\%$  confidence interval for  $\mu_2 - \mu_1$  is

$$CI_{1-\alpha} = \bar{X}_2 - \bar{X}_1 \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_2} + \frac{\sigma_2^2}{n_2}}.$$

**Example 14.4** A sample of 50 engineers from Italy is selected, and it is found that the average salary was 56,023. An additional sample of 42 engineers from France is selected, who have among them an average salary of 59,587. Suppose from a previous study it is known that the population variances of engineer's salaries from Italy and France are 23,985 and 25,487 respectively.

To construct a 95% confidence interval, letting Italian engineers be population 1, we make note that

$$\begin{aligned} n_1 &= 50 \\ n_2 &= 42 \\ \bar{X}_1 &= 56,023 \\ \bar{X}_2 &= 59,587 \\ \sigma_1^2 &= 23,985 \\ \sigma_2^2 &= 25,487 \\ \alpha &= .05 \\ z_{.025} &= 1.96 \end{aligned}$$

Then making the appropriate substitutions gives

$$\begin{aligned} CI_{1-\alpha} &= \bar{X}_2 - \bar{X}_1 \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \\ &= 59,587 - 56,023 \pm 1.96 \sqrt{\frac{23,985}{50} + \frac{25,487}{42}} \\ &= 3,564 \pm 1.96 \sqrt{479.7 + 606.8} \\ &= 3,564 \pm 64.6 \end{aligned}$$

Note that the confidence interval does not contain zero. This gives evidence that French salaries are higher than Italian salaries. ■

For testing statistical hypotheses we use the test statistic

$$Z_{obs} = \frac{\bar{X}_2 - \bar{X}_1 - \Delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

Given this test statistic, the observed significance levels are calculated according to the following rules.

1. One sided, lower tailed hypothesis

$$\alpha_{obs} = P(Z < Z_{obs})$$

2. One sided, upper tailed hypothesis

$$\alpha_{obs} = P(Z > Z_{obs})$$

3. Two sided hypothesis

$$\begin{aligned} \alpha_{obs} &= P(Z < -|Z_{obs}| \text{ O } |Z_{obs}| < Z) \\ &= 2P(Z > |Z_{obs}|) \end{aligned}$$

where we assume  $Z \sim N(0, 1)$ .

There will be three different forms of the rejection region depending on the form of the hypothesis.

1. One sided, lower tailed hypothesis

$$\text{Reject } H_o \text{ if } Z_{obs} \leq -z_\alpha$$

2. One sided, upper tailed hypothesis

$$\text{Reject } H_o \text{ if } Z_{obs} \geq z_\alpha$$

3. Two sided hypothesis

$$\text{Reject } H_o \text{ if } |Z_{obs}| \geq z_\alpha/2$$

where  $z_\alpha$  is the  $\alpha$  critical value from a standard normal distribution.

**Example 14.5** We will treat the previous example as a statistical hypothesis, where we are interested in testing to see if there is a difference in salaries between French and Italian engineers.

Since this is a two-sided hypothesis we have

$$\begin{aligned} H_o &: \mu_2 = \mu_1 \\ H_a &: \mu_2 \neq \mu_1 \end{aligned}$$

where we designate the Italian population as population 1. The hypotheses can be written this way, since  $\Delta = 0$ .

The test statistic becomes

$$\begin{aligned} Z_{obs} &= \frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \\ &= \frac{59,587 - 56,023}{\sqrt{\frac{23,985}{50} + \frac{25,487}{42}}} \\ &= \frac{3,564}{\sqrt{479.7 + 606.8}} \\ &= \frac{3,564}{33.0} \\ &= 108.0. \end{aligned}$$

The observed test statistic is well beyond the range in which we need consult the standard normal distribution tables, so we may conclude

$$\alpha_{obs} < 0.0002$$

since 0.0002 is the smallest entry in the table. Therefore, there is strong evidence that the French engineer's salaries differ from Italian engineer's salaries. ■

### 14.1.2 Case 2 - Variances are Unknown but Equal

If the population variances are unknown, we rely on the sample variances to estimate them. If we know that the population variances are equal, then it makes sense to combine all the data to form one estimate of the single (but unknown) population variance. We may calculate a *pooled variance* to serve as this estimate.

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 2)S_2^2}{n_1 + n_2 - 2}$$

Then for this case, the  $(1 - \alpha) * 100\%$  confidence interval for  $\mu_2 - \mu_1$  is

$$CI_{1-\alpha} = \bar{X}_2 - \bar{X}_1 \pm t_{n_1+n_2-2, \alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Note that the confidence interval is based on a *t*-distribution with  $n_1 + n_2 - 2$  degrees of freedom.

**Example 14.6** The Chapin Social Insight test measures how accurately the subject appraises other people. Possible scores range from 0 to 41. The test was applied to independent samples of males and females with the following results

$i$	Pop'n	$n_i$	$\bar{X}_i$	$S_i$
1	Male	13	25.34	5.05
2	Female	16	24.94	5.44

Suppose it is known from previous experience with this test that the variance does not differ significantly between male and female subjects, so that we may assume that

$$\sigma_1^2 = \sigma_2^2.$$

We therefore calculate the pooled variance

$$\begin{aligned} S_p^2 &= \frac{(n_1 - 1)S_1^2 + (n_2 - 2)S_2^2}{n_1 + n_2 - 2} \\ &= \frac{(13 - 1)5.05^2 + (16 - 1)5.44^2}{13 + 16 - 2} \\ &= \frac{749.9}{27} \\ &= 27.8. \end{aligned}$$

To calculate a 98% confidence interval we have  $\alpha = 0.02$  so we require the .02/2 critical value of a *t*-distribution with  $n_1 + n_2 - 2 = 27$  degrees of freedom. Directly from the table we get

$$t_{27, 0.01} = 2.4727$$

Then, substituting into the formula for the confidence interval we get

$$\begin{aligned} CI_{1-\alpha} &= \bar{X}_2 - \bar{X}_1 \pm t_{n_1+n_2-2,\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\ &= 24.94 - 25.34 \pm 2.4727 \sqrt{27.8} \sqrt{\frac{1}{13} + \frac{1}{16}} \\ &= -0.4 \pm 4.868 \end{aligned}$$

Note that the confidence interval contains 0. This means that 0 is a plausible value for  $\mu_2 - \mu_1$ , so that we could not reject this a possibility. ■

For testing statistical hypotheses we use the test statistic

$$T_{obs} = \frac{\bar{X}_2 - \bar{X}_1 - \Delta}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

Given this test statistic, the observed significance levels are calculated according to the following rules.

1. One sided, lower tailed hypothesis

$$\alpha_{obs} = P(T < T_{obs})$$

2. One sided, upper tailed hypothesis

$$\alpha_{obs} = P(T > T_{obs})$$

3. Two sided hypothesis

$$\begin{aligned} \alpha_{obs} &= P(T < -|T_{obs}| \text{ OR } |T_{obs}| < T) \\ &= 2P(T > |T_{obs}|) \end{aligned}$$

where we assume  $T$  has a  $t$ -distribution with  $n_1 + n_2 - 2$  degrees of freedom.

There will be three different forms of the rejection region depending on the form of the hypothesis.

1. One sided, lower tailed hypothesis

Reject  $H_o$  if  $T_{obs} \leq -t_{n_1+n_2-2,\alpha}$

2. One sided, upper tailed hypothesis

Reject  $H_o$  if  $T_{obs} \geq t_{n_1+n_2-2,\alpha}$

### 3. Two sided hypothesis

Reject  $H_o$  if  $|T_{obs}| \geq t_{n_1+n_2-2,\alpha/2}$

where  $t_{n_1+n_2-2,\alpha}$  is the  $\alpha$  critical value from a  $t$ -distribution with  $n_1 + n_2 - 2$  degrees of freedom.

**Example 14.7** We will treat the previous example as a statistical hypothesis. We suppose that it was conjectured that male subjects score higher on average than female subjects. Since this is a one sided, lower tailed hypothesis we have

$$H_o : \mu_2 \geq \mu_1$$

$$H_a : \mu_2 < \mu_1$$

where we designate the male population as population 1. Here,  $\Delta = 0$ .

The test statistic becomes

$$\begin{aligned} T_{obs} &= \frac{\bar{X}_2 - \bar{X}_1}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\ &= \frac{24.94 - 25.34}{\sqrt{27.8} \sqrt{\frac{1}{13} + \frac{1}{16}}} \\ &= -0.203 \end{aligned}$$

The number degrees of freedom is  $n_1 + n_2 - 2 = 27$ . From the tables we have

$$t_{27,0.25} = 0.6837.$$

Since

$$t_{27,0.25} > |T_{obs}|$$

we may conclude that

$$\alpha_{obs} > 0.25$$

so that there is no evidence that male test scores are higher on average than female test scores. ■

#### 14.1.3 Case 3 - Variances are Unknown and Not Assumed Equal

If the variances are unknown, and no other assumptions are to be made, then we may proceed as if we do know the variances, substituting the sample variances for them. The adjustment we make is in the critical values. Instead of using a normal distribution critical value we use a  $t$ -distribution with an

*estimated* degrees of freedom. One simple approach is to use the conservative estimate:

$$\nu_{min} = \min(n_1 - 1, n_2 - 1)$$

This has a tendency to overcompensate for the additional variation due to the use of the sample variances. A more accurate approximation is given by the *Welch t-test*. This can also be referred to as *Satterthwaite approximation*, a more general approximation method on which Welch's t-test is based. It is also referred to as the *Smith-Satterthwaite* procedure, acknowledging preceding work. It is given by

$$\nu_W = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}}.$$

This quantity will generally not be an integer, so it will be appropriate to round down.

In this case, the  $(1 - \alpha)100\%$  confidence interval for  $\mu_2 - \mu_1$  is

$$CI_{1-\alpha} = \bar{X}_2 - \bar{X}_1 \pm t_{\nu, \alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}},$$

where  $\nu$  is the estimated degrees of freedom.

**Example 14.8** Suppose a lake was restocked with trout. A sample of size 20 was taken before the restocking, then a sample of size 26 was taken afterwards. It is conjectured that the restocking has the effect of increasing the average weight. The data obtained are summarized below (in pounds)

$i$	Pop'n	$n_i$	$\bar{X}_i$	$S_i$
1	Before restock	20	1.57	0.32
2	After restock	26	1.86	0.45

To construct a 95% confidence interval, letting the before-restock lake

be population 1, we note that

$$\begin{aligned}
 n_1 &= 20 \\
 n_2 &= 26 \\
 \nu_{min} &= \min(n_1 - 1, n_2 - 1) \\
 &= \min(19, 25) \\
 &= 19 \\
 \nu_W &= \frac{\left(\frac{0.32^2}{20} + \frac{0.45^2}{26}\right)^2}{\frac{(0.32^2/20)^2}{20-1} + \frac{(0.45^2/26)^2}{26-1}} \\
 &\approx 43.78 \\
 \alpha &= .05 \\
 t_{19,.025} &= 2.093
 \end{aligned}$$

Then making the appropriate substitutions gives

$$\begin{aligned}
 CI_{1-\alpha} &= \bar{X}_2 - \bar{X}_1 \pm t_{19,\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \\
 &= 1.86 - 1.57 \pm 2.093 \sqrt{\frac{0.32^2}{20} + \frac{0.45^2}{26}} \\
 &= 0.29 \pm 0.238
 \end{aligned}$$

using estimate  $\mu_{min} = 19$  or

$$\begin{aligned}
 CI_{1-\alpha} &= \bar{X}_2 - \bar{X}_1 \pm t_{43,\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \\
 &= 1.86 - 1.57 \pm 2.017 \sqrt{\frac{0.32^2}{20} + \frac{0.45^2}{26}} \\
 &= 0.29 \pm 0.229
 \end{aligned}$$

using the Welch t-test, based on estimate  $\nu_W = 43$  (after rounding down). Note this is very close to the degrees of freedom appropriate for the pooled t-test, that is,  $n_1 + n_2 - 2 = 44$ . While the Welch procedure gives a narrower confidence interval, the difference is not great. ■

For testing statistical hypotheses we use the test statistic

$$T_{obs} = \frac{\bar{X}_2 - \bar{X}_1 - \Delta}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}.$$

Given this test statistic, the observed significance levels are calculated according to the following rules.

1. One sided, lower tailed hypothesis

$$\alpha_{obs} = P(T < T_{obs})$$

2. One sided, upper tailed hypothesis

$$\alpha_{obs} = P(T > T_{obs})$$

3. Two sided hypothesis

$$\begin{aligned}\alpha_{obs} &= P(T < -|T_{obs}| \text{ OR } |T_{obs}| < T) \\ &= 2P(T > |T_{obs}|)\end{aligned}$$

where we assume  $T$  has a  $t$ -distribution with  $\nu$  degrees of freedom.

There will be three different forms of the rejection region depending on the form of the hypothesis.

1. One sided, lower tailed hypothesis

$$\text{Reject } H_o \text{ if } T_{obs} \geq -t_{\nu,\alpha}$$

2. One sided, upper tailed hypothesis

$$\text{Reject } H_o \text{ if } T_{obs} \leq t_{\nu,\alpha}$$

3. Two sided hypothesis

$$\text{Reject } H_o \text{ if } |T_{obs}| \geq t_{\nu,\alpha/2}$$

where  $t_{\nu,\alpha}$  is the  $\alpha$  critical value from a  $t$ -distribution with  $\nu$  degrees of freedom.

**Example 14.9** It is conjectured that the restocking has the effect of increasing the average weight.

Since this is a one sided upper tailed hypothesis we have

$$\begin{aligned}H_o &: \mu_2 \leq \mu_1 \\ H_a &: \mu_2 > \mu_1\end{aligned}$$

where we designate the pre-restocked population as population 1, with  $\Delta = 0$ .

The test statistic becomes

$$\begin{aligned} T_{obs} &= \frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \\ &= \frac{1.86 - 1.57}{\sqrt{\frac{0.32^2}{20} + \frac{0.45^2}{26}}} \\ &= 2.55 \end{aligned}$$

We use a  $t$ -distribution with  $\nu_{min} = 19$  degrees of freedom. From the tables we have

$$\begin{aligned} t_{19,0.01} &= 2.539 \\ &< T_{obs} \\ &< 2.861 \\ &= t_{19,0.005} \end{aligned}$$

from which we conclude

$$0.005 < \alpha_{obs} < 0.01$$

so that there is strong evidence of an increase in average weight. ■

## 14.2 Paired Samples

If a single sample from a population is collected, and two distinct types of measurements are taken from each unit (person, family, etc.) sampled, we then have two samples that are *paired*. The resulting data set has the following structure.

Sampling Unit	Measurement Type 1	Measurement Type 2	Difference
1	$X_1$	$Y_1$	$D_1 = Y_1 - X_1$
2	$X_2$	$Y_2$	$D_2 = Y_2 - X_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n$	$X_n$	$Y_n$	$D_n = Y_n - X_n$

Here, we sampled  $n$  units. From the  $i$ th units two kinds on measurements  $X_i$  and  $Y_i$  were made. We assume that for each kind of measurement there is a population mean calculable from all units in the population. If interest

is in the difference between the two populations means, then for each pair of measurements we calculate

$$D_i = Y_i - X_i$$

for each observation pair.

We label the two population means  $\mu_1$  and  $\mu_2$ . We assume interest is in  $\mu_2 - \mu_1$ . Note that if the population mean from which the  $X_i$ 's are sampled is  $\mu_1$  and the population mean from which the  $Y_i$ 's are sampled is  $\mu_2$  then we can consider the observations  $D_i$  to be a single sample from a population with mean  $\mu_2 - \mu_1$ . Accordingly, we can use techniques appropriate for inference about a single population mean based on a single sample, where we assume

$$D_1, D_2, \dots, D_n$$

is the sample of size  $n$  and  $\mu_2 - \mu_1$  is the population mean. We therefore set

$$\bar{D} = \frac{\sum_{i=1}^n D_i}{n}$$

as the sample mean, and the sample variance is

$$\begin{aligned} S_D^2 &= \frac{\sum_{i=1}^n (D_i - \bar{D})^2}{n-1} \\ &= \frac{\sum_{i=1}^n D_i^2 - n\bar{D}^2}{n-1} \\ &= \frac{\sum_{i=1}^n D_i^2 - \frac{(\sum_{i=1}^n D_i)^2}{n}}{n-1}. \end{aligned}$$

A level  $(1 - \alpha)100\%$  confidence interval for  $\mu_2 - \mu_1$  is given by

$$CI_{1-\alpha} = \bar{D} \pm t_{n-1, \alpha/2} \frac{S_D}{\sqrt{n}},$$

which is based on the  $t$ -distribution with  $n - 1$  degrees of freedom.

**Example 14.10** This example is from Maskin *et al.*, 1985 in the journal *Circulation*. The following table shows the measurement of heart rate (in beats per minute) before and 30 minutes after administering enalprilat to 9 subjects.

Subject	Time after treatment 0 mins	Time after treatment 30 mins	Change in heart rate
1	96	92	-4
2	110	106	-4
3	89	86	-3
4	95	78	-17
5	128	124	-4
6	100	98	-2
7	72	68	-4
8	79	75	-4
9	100	106	+6

Suppose we let  $\mu_1$  be the mean pretreatment heart rate among all eligible subjects in the population, and let  $\mu_2$  be the corresponding post-treatment heart rate. To construct a 95% confidence interval for  $\mu_2 - \mu_1$  we first note that this is a paired sample, so we calculate the appropriate differences, and use this as the sample. From the above table we have

$$\begin{aligned} D_1 &= -4 \\ D_2 &= -4 \\ &\vdots \\ D_9 &= +6 \end{aligned}$$

giving

$$\begin{aligned} \bar{D} &= -4 \\ S_D^2 &= 34.25 \end{aligned}$$

Since  $n - 1 = 8$  we have from the tables

$$t_{8,0.025} = 2.306$$

so that

$$\begin{aligned} CI_{95\%} &= \bar{D} \pm t_{n-1,\alpha/2} \frac{S_D}{\sqrt{n}} \\ &= -4 \pm 2.306 \sqrt{\frac{34.25}{9}} \\ &= -4 \pm 4.77 \end{aligned}$$

gives the 95% confidence interval. ■

For testing statistical hypotheses we use the test statistic

$$T_{obs} = \frac{\bar{D} - \Delta}{S_D / \sqrt{n}}.$$

Given this test statistic, the observed significance levels are calculated according to the following rules.

1. One sided, lower tailed hypothesis

$$\alpha_{obs} = P(T < T_{obs})$$

2. One sided, upper tailed hypothesis

$$\alpha_{obs} = P(T > T_{obs})$$

3. Two sided hypothesis

$$\begin{aligned}\alpha_{obs} &= P(T < -|T_{obs}| \text{ OR } |T_{obs}| < T) \\ &= 2P(T > |T_{obs}|)\end{aligned}$$

where we assume  $T$  has a  $t$ -distribution with  $n - 1$  degrees of freedom.

There will be three different forms of the rejection region depending on the form of the hypothesis.

1. One sided, lower tailed hypothesis

Reject  $H_o$  if  $T_{obs} \leq -t_{n-1,\alpha}$

2. One sided, upper tailed hypothesis

Reject  $H_o$  if  $T_{obs} \geq t_{n-1,\alpha}$

3. Two sided hypothesis

Reject  $H_o$  if  $|T_{obs}| \geq t_{n-1,\alpha/2}$

where  $t_{n-1,\alpha}$  is the  $\alpha$  critical value from a  $t$ -distribution with  $n - 1$  degrees of freedom.

**Example 14.11** We will treat the previous example as a statistical hypothesis, where we are interested in testing to see if there is a decrease in heart rate after the treatment is administered.

Since this is a lower tailed hypothesis we have

$$\begin{aligned} H_o &: \mu_2 \geq \mu_1 \\ H_a &: \mu_2 < \mu_1 \end{aligned}$$

where we designate the pre-treatment measure population as population 1, with  $\Delta = 0$ .

The test statistic becomes

$$\begin{aligned} T_{obs} &= \frac{\bar{D}}{S_D/\sqrt{n}} \\ &= \frac{-4}{\sqrt{\frac{34.25}{9}}} \\ &= -2.05 \end{aligned}$$

To calculate the observed significance level we use, assuming that  $T$  has a  $t$ -distribution with  $n - 1 = 8$  degrees of freedom

$$\begin{aligned} \alpha_{obs} &= P(T < T_{obs}) \\ &= P(T < -2.05) \\ &= P(T > 2.05). \end{aligned}$$

From the  $t$ -distribution tables we have

$$\begin{aligned} t_{8,0.05} &= 1.8331 \\ t_{8,0.025} &= 2.2622 \end{aligned}$$

from which we conclude

$$0.025 < \alpha_{obs} < 0.05.$$

■

### 14.3 The $t$ -test in R

The function `t.test()` is used for all types of  $t$ -test in R. The options are:

```
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, ...)
```

Data is entered as a vector `x` for a one sample procedure, and as two vectors `x,y` for a two sample procedure. The hypothetical mean, or difference in mean, as appropriated is specified by option `mu`, with default 0, the usual choice for a two-sample test. The test may be paired. For a two sample test, setting `var.equal` to `FALSE` gives the Welch procedure, otherwise the pooled procedure is used. A confidence interval is also given.

Consider the following script:

```
> x = rnorm(100, mean = 10, sd = 1)
> t.test(x, mu = 8)

One Sample t-test

data: x
t = 18.7734, df = 99, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 8
95 percent confidence interval:
 9.809908 10.237712
sample estimates:
mean of x
10.02381
```

```
> y = rnorm(50, mean = 10, sd = 2)
> t.test(x,y)
```

Welch Two Sample t-test

```
data: x and y
t = 0.2742, df = 64.102, p-value = 0.7848
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.5142560 0.6778673
sample estimates:
mean of x mean of y
10.023810 9.942004

>
> t.obj = t.test(x,y)
> names(t.obj)
[1] "statistic"    "parameter"    "p.value"      "conf.int"
"estimate"      "null.value"   "alternative"  "method"
"data.name"
> t.obj$p.value
[1] 0.7848377
> t.obj$statistic
t
0.2741672
> t.obj$conf.int
[1] -0.5142560 0.6778673
attr(,"conf.level")
```

```
[1] 0.95
>
```

Note that the output of an R function is often a list, which can be stored as another object. In the above example, this is accomplished by the command `> t.obj = t.test(x,y)`. To see the object labels, use the command `names(t.obj)`. In this way, the specific quantities associated with the t-test, such as the p-value or the confidence interval, can be captured and stored as an variable or object for subsequent use.

Also note that the appropriate quantile function can be used to obtain critical values, for example:

```
> qt(1 - 0.025, df = 43)
[1] 2.016692
```

gives  $t_{43,0.025} \approx 2.017$  in Example 14.8.

## 14.4 Assumptions

The assumptions required by the techniques of this section, besides those explicitly stated, are that the sample is a true random sample, and that the underlying population distribution is normally distributed. In practice, inference using the *t*-distribution is reasonably accurate when the data is not too skewed. Note that in the case of the paired samples the assumptions are applicable to the differences and not to the original two samples.

## Chapter 15

# Inference for Population Proportions

### 15.1 Single Population Proportion

Sometimes we are interested in the frequency with which a *type* of unit occurs in a population. A common example is the opinion poll, in which it is assumed that there is a fixed proportion of the population who has a given opinion on a certain matter.

Suppose the proportion in a population of a certain type is  $p$ . To estimate  $p$  we take a random sample of size  $n$  from the population. If  $\hat{p}$  is the proportion in the sample of the type of interest, then this serves as an estimate of  $p$ . Furthermore, a consequence of the Central Limit Theorem is that, approximately

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right).$$

This means that the standard deviation of  $\hat{p}$  is

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}.$$

Of course, if we are trying to estimate  $p$  that means we don't know its' value. We can, however, approximate the standard deviation by substituting  $\hat{p}$  for  $p$ , to get

$$\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

This leads to a level  $(1 - \alpha)100\%$  confidence interval for  $p$  given by

$$CI_{\alpha/2} = \hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

**Example 15.1** An recent Gallup opinion poll taken of 1,022 adults asked the question "Do you think the outcome of the air strikes in Kosovo represents a victory for the United States, or not?" Of those asked, 40% replied "Yes".

To construct a 95% confidence interval we note that

$$n = 1,022$$

$$\hat{p} = 0.4$$

making the appropriate substitutions gives the confidence interval

$$\begin{aligned} CI_{95\%} &= \hat{p} \pm z_{0.025} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \\ &= 0.40 \pm 1.96 \sqrt{\frac{0.4(1 - 0.4)}{1022}} \\ &= 0.40 \pm 0.03. \end{aligned}$$

In the usual terminology of surveys we say that the estimate of the percentage of those who answer "yes" is 40% with a margin of error of 3% 19 times out of 20 ( $19/20 = 95\%$ ). ■

For hypothesis testing we again have three types of hypotheses. Suppose we set  $p_0$  to be some hypothetical population proportion, and let  $p$  be the true population proportion.

1. One sided, lower tailed hypothesis

$$H_o : p \geq p_0$$

$$H_a : p < p_0$$

We are looking for evidence that  $p$  is **less than**  $p_0$ .

2. One sided, upper tailed hypothesis

$$H_o : p \leq p_0$$

$$H_a : p > p_0$$

We are looking for evidence that  $p$  is **greater than**  $p_0$ .

3. Two sided hypothesis

$$H_o : p = p_0$$

$$H_a : p \neq p_0$$

We are looking for evidence that  $p$  is **not equal to**  $p_0$ .

We then have the test statistic

$$Z_{obs} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

Note that we use the hypothetical value  $p_0$  in the test statistic.

Given this test statistic, the observed significance levels are calculated according to the following rules.

1. One sided, lower tailed hypothesis

$$\alpha_{obs} = P(Z < Z_{obs})$$

2. One sided, upper tailed hypothesis

$$\alpha_{obs} = P(Z > Z_{obs})$$

3. Two sided hypothesis

$$\begin{aligned}\alpha_{obs} &= P(Z < -|Z_{obs}| \text{ OR } |Z_{obs}| < Z) \\ &= 2P(Z > |Z_{obs}|)\end{aligned}$$

where we assume  $Z \sim N(0, 1)$ .

There will be three different forms of the rejection region depending on the form of the hypothesis.

1. One sided, lower tailed hypothesis

Reject  $H_o$  if  $Z_{obs} \leq -z_\alpha$

2. One sided, upper tailed hypothesis

Reject  $H_o$  if  $Z_{obs} \geq z_\alpha$

3. Two sided hypothesis

Reject  $H_o$  if  $|Z_{obs}| \geq z_{\alpha/2}$

where  $z_\alpha$  is the  $\alpha$  critical value from a standard normal distribution.

**Example 15.2** We continue the previous example. Suppose the newspaper which commissioned the poll is interested particularly in the conjecture that less than a majority believe the air strikes were a victory. In this case the appropriate hypothesis is

$$H_o : p \geq 0.5$$

$$H_a : p < 0.5$$

where we have hypothetical value  $p_0 = 0.5$ . The test statistic is then

$$\begin{aligned} Z_{obs} &= \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \\ &= \frac{0.4 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{1022}}} \\ &= \frac{-0.1}{0.0156} \\ &= -6.41 \end{aligned}$$

We note that this value is below the lower limit of the standard normal tables, so we may conclude

$$\alpha_{obs} < 0.0002$$

so that there is strong evidence that less than a majority would answer "yes". ■

## 15.2 Difference Between Two Population Proportions

Sometimes interest will be in the difference between the proportions of a certain category in two different populations. For example, the proportion of the voting population which supports a certain political party may differ between two provinces. In this case we may take a random sample from each province, estimate the proportions within each province, then examine the difference between these two estimates.

Assume that the population proportions of interest are  $p_1$  and  $p_2$ , and that random samples of size  $n_1$  and  $n_2$  are selected from each. Furthermore, suppose that the proportions of interest observed in the two samples are  $\hat{p}_1$  and  $\hat{p}_2$ . We further assume that the two samples were collected independently. Then a level  $(1 - \alpha)100\%$  confidence interval for  $p_2 - p_1$  is given by

$$CI_{1-\alpha} = \hat{p}_2 - \hat{p}_1 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}.$$

**Example 15.3** Two surveys were conducted which asked those who purchased a certain brand of computer whether they were likely to purchase the same brand next time they bought a computer. The first survey was conducted in June of 1998, in which 1,000 respondents were surveyed. Of these, 42% said they would be likely to purchase the same brand. The second survey was conducted in January of 1999, in which 1,500 respondents

were surveyed. Of these, 45% said they would be likely to purchase the sample brand.

To construct a 95% confidence interval, we have

$$\begin{aligned} n_1 &= 1,000 \\ n_2 &= 1,500 \\ \hat{p}_1 &= 0.42 \\ \hat{p}_2 &= 0.45 \\ \alpha &= 0.05 \\ z_{\alpha/2} &= 1.96 \end{aligned}$$

giving as a confidence interval

$$\begin{aligned} CI_{.95} &= \hat{p}_2 - \hat{p}_1 \pm z_{0.025} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \\ &= 0.45 - 0.42 \pm \sqrt{\frac{0.42(1 - 0.42)}{1,000} + \frac{0.45(1 - 0.45)}{1,500}} \\ &= 0.03 \pm 1.96 \times 0.0202 \\ &= 0.03 \pm 0.04 \\ &= (-0.00962, 0.0696) \end{aligned}$$

so that the estimated change in proportion is 3% with a margin of error of 4% with a confidence level of 95%. ■

For hypothesis testing we again have three types of hypotheses. Suppose we set  $p_1$  and  $p_2$  to be the two population proportions. We then have the following hypotheses.

1. One sided, lower tailed hypothesis

$$\begin{aligned} H_o : p_2 &\geq p_1 \\ H_a : p_2 &< p_1 \end{aligned}$$

We are looking for evidence that  $p_2$  is **less than**  $p_1$ .

2. One sided, upper tailed hypothesis

$$\begin{aligned} H_o : p_2 &\leq p_1 \\ H_a : p_2 &> p_1 \end{aligned}$$

We are looking for evidence that  $p_2$  is **greater than**  $p_1$ .

3. Two sided hypothesis

$$\begin{aligned} H_o &: p_2 = p_1 \\ H_a &: p_2 \neq p_1 \end{aligned}$$

We are looking for evidence that  $p_2$  is **not equal to**  $p_1$ .

Note that for the null hypothesis we may set  $p_0 = p_1 = p_2$ . Since we usually construct the test statistic to have a certain distribution assuming that  $H_o$  is true, it makes sense to combine the two samples to form a single estimate of  $p_0$ , referred to as a **pooled** estimate of  $p_0$ . This is given by

$$\hat{p}_0 = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

We then have the test statistic

$$Z_{obs} = \frac{\hat{p}_2 - \hat{p}_1}{\sqrt{\hat{p}_0(1 - \hat{p}_0) \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Note that we use the pooled estimate  $\hat{p}_0$  in the test statistic.

Given this test statistic, the observed significance levels are calculated according to the following rules.

1. One sided, lower tailed hypothesis

$$\alpha_{obs} = P(Z < Z_{obs})$$

2. One sided, upper tailed hypothesis

$$\alpha_{obs} = P(Z > Z_{obs})$$

3. Two sided hypothesis

$$\begin{aligned} \alpha_{obs} &= P(Z < -|Z_{obs}| \text{ OR } |Z_{obs}| < Z) \\ &= 2P(Z > |Z_{obs}|) \end{aligned}$$

where we assume  $Z \sim N(0, 1)$ .

There will be three different forms of the rejection region depending on the form of the hypothesis.

1. One sided, lower tailed hypothesis

Reject  $H_o$  if  $Z_{obs} \leq -z_\alpha$

2. One sided, upper tailed hypothesis

$$\text{Reject } H_o \text{ if } Z_{obs} \geq z_\alpha$$

3. Two sided hypothesis

$$\text{Reject } H_o \text{ if } |Z_{obs}| \geq z_{\alpha/2}$$

where  $z_\alpha$  is the  $\alpha$  critical value from a standard normal distribution.

**Example 15.4** We continue the previous example. Suppose after the first survey, the company implemented improvements in its product, in the hopes of increasing the positive response of the survey. Is there evidence that the positive response has increased?

This would be an upper-tailed test, using hypotheses

$$H_o : p_2 \leq p_1$$

$$H_a : p_2 > p_1.$$

The pooled estimate of  $p_0$  is

$$\begin{aligned}\hat{p}_0 &= \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} \\ &= \frac{1,000 \times 0.42 + 1,500 \times 0.45}{1,000 + 1,500} \\ &= 0.438\end{aligned}$$

The test statistic is then

$$\begin{aligned}Z_{obs} &= \frac{\hat{p}_2 - \hat{p}_1}{\sqrt{\hat{p}_0(1 - \hat{p}_0) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \\ &= \frac{0.45 - 0.42}{\sqrt{0.438(1 - 0.438) \left( \frac{1}{1,000} + \frac{1}{1,500} \right)}} \\ &= 1.48\end{aligned}$$

If we test at a significance level of 5%, we reject  $H_o$  if  $Z_{obs} \geq z_{0.05} = 1.645$ . Since  $Z_{obs}$  does not fall in the rejection region, then there is not sufficient evidence to reject the hypothesis that the proportion of positives did not increase between the two surveys. ■

### 15.3 Binomial Continuity Correction

If  $X_{bin} \sim bin(n, p)$  and  $X_{norm} \sim N(np, np(1-p))$  then we have approximation

$$P(X_{bin} = k) \approx P(k - 0.5 \leq X_{norm} \leq k + 0.5).$$

This means the CDF of  $X_{bin}$  should use the approximation

$$F_{X_{bin}}(k) = P(X_{bin} \leq k) \approx P(X_{norm} \leq k + 0.5) = F_{X_{norm}}(k + 0.5). \quad (15.1)$$

### 15.4 Inference for the Odds Ratio

The odds ratio was introduced in Section 5.5 as a method of comparing proportions, and we return to the discussion. Consider the following events

$$\begin{aligned} O_- &= \{ \text{negative outcome} \} \\ O_+ &= \{ \text{positive outcome} \} \\ G_1 &= \{ \text{the patient is in Group 1} \} \\ G_2 &= \{ \text{the patient is in Group 2} \}. \end{aligned}$$

The outcome may be *infection* or *cancer recurrence*. The groups are chosen with various possible intentions. These include discovery of risk factors for the outcome, or demonstration of the efficacy of a treatment. In the latter case,  $G_1$  and  $G_2$  will be two treatments. Often, one of these is a control, which is essentially no treatment (this may involve a *placebo*, or *sham* treatment). In a *randomized clinical trial* subjects from a homogeneous pool are randomly assigned to the two treatment groups, resulting in two samples. The importance of this is that from a statistical point of view, the only difference between the two samples is in the treatments, and so any difference in outcomes must be attributable to the difference in treatment.

Another type of study is the *case-control* study. In this type of study, subjects are identified by outcome, then differences in risk factors are analyzed. This type of study is generally easier to carry out, and is especially appropriate when the prevalence of a positive outcome is small. However, it is more difficult to establish the type of causal relationship that a randomized clinical trial is designed to establish. In addition, there is less information about the actual magnitude of risk, since the numbers of positive and negative outcomes are determined in advance.

In statistical terms we are interested in comparing

$$P(O_+ | G_1) \text{ and } P(O_+ | G_2).$$

We have already seen methods of comparing two population proportions by examining their difference. However, when the probabilities are small,

this method may not work well. In addition, for a case-control study, this difference would not be interpretable.

Alternatively, we have the *relative risk*

$$RR = \frac{P(O_+ | G_1)}{P(O_+ | G_2)}.$$

This quantity is more suitable for small probabilities, but is also not interpretable for a case-control study.

A frequently used quantity is the *odds ratio*

$$OR = \frac{Odds(O_+ | G_1)}{Odds(O_+ | G_2)} = \frac{P(O_+ | G_1)/(1 - P(O_+ | G_1))}{P(O_+ | G_2)/(1 - P(O_+ | G_2))}.$$

The events defining the OR may be transposed, that is

$$OR = \frac{Odds(G_1 | O_+)}{Odds(G_1 | O_-)} = \frac{Odds(O_+ | G_1)}{Odds(O_+ | G_2)}.$$

Essentially, the odds ratio does not depend on the prevalence  $P(O_+)$ , so that the OR may be used for case-control studies, and is also suitable for comparing small probabilities.

Statistically, the OR can proceed by using a  $2 \times 2$  contingency table:

	$O_+$	$O_-$	
$G_1$	$n_{11}$	$n_{12}$	$R_1$
$G_2$	$n_{21}$	$n_{22}$	$R_2$
Total	$C_1$	$C_2$	$n$

The formula for the odds ratio  $OR$  is simply

$$OR = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

Inference proceeds by a natural log transformation  $\log(OR)$  of  $OR$ , which has standard error

$$SE(\log(OR)) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}.$$

We accept a large sample normal approximation of  $\log(OR)$ .

**Example 15.5** For table

	$O_+$	$O_-$	
$G_1$	32	118	150
$G_2$	17	127	144
Total	49	245	294

we have

$$OR = \frac{32 \times 127}{17 \times 118} = 2.026$$

and

$$SE(\log(OR)) = \sqrt{1/32 + 1/118 + 1/17 + 1/127} = 0.326,$$

with  $\log(OR) = 0.706$ . This gives a 95% CI for the log odds ratio

$$0.706 \pm 1.96 \times 0.326 = 0.706 \pm 0.64.$$

This means we can reject the null hypothesis  $H_0 : OR = 1$  (equivalently  $H_0 : \log(OR) = 0$ ). Since  $OR > 1$  there is a positive association between  $G_1$  and  $O_+$ . ■

## 15.5 Testing for Proportions in R

Tests for proportions (1 and 2 samples) are supported in R using the function `prop.test()`. The options are as follows

```
prop.test(x, n, p = NULL,
          alternative = c("two.sided", "less", "greater"),
          conf.level = 0.95, correct = TRUE)
```

Sample proportions are entered using a count numerator `x` and sample size denominator `n`. These objects are vectors containing one element for each sample. More than 2 samples may be used. The object `p` defines the null hypothesis, for multiple as well as single samples. If `p` is not specified, then for the one sample case a null hypothesis of  $p = 1/2$  is assumed, and for multiple sample the null hypothesis is the equality of proportions.

Example 15.3 may be calculated in the following way:

```
> x1 = 0.42*1000
> x2 = 0.45*1500
> prop.test(x = c(x2, x1), n = c(1500,1000), correct=F)
```

2-sample test for equality of proportions without continuity correction

```
data: c(x2, x1) out of c(1500, 1000)
X-squared = 2.1937, df = 1, p-value = 0.1386
alternative hypothesis: two.sided
95 percent confidence interval:
-0.009618431 0.069618431
sample estimates:
prop 1 prop 2
0.45 0.42
```

**Example 15.6** In Diaconis *et al.* (2007) the Stanford probabilist Persi Diaconis and colleagues argued that there is a bias in coin tossing in favor of the side facing up at the start of the toss. The biased proportion was estimated to be 50.8% in place of the commonly expected 50%. In 2009 two Berkeley students each flipped a coin 20,000 times, one starting with Heads facing up, the other with Tails facing up. See [http://www.stat.berkeley.edu/~aldous/Real-World/coin\\_tosses.html](http://www.stat.berkeley.edu/~aldous/Real-World/coin_tosses.html) for details.

The two students attained 10231 Heads, and 10014 Tails according to the respective starting conditions. This leads to the following estimates, given separately and combined:

```
> prop.test(x = 10014, n = 20000)

1-sample proportions test with continuity correction

data: 10014 out of 20000, null probability 0.5
X-squared = 0.0365, df = 1, p-value = 0.8486
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4937460 0.5076537
sample estimates:
      p
0.5007

> prop.test(x = 10231, n = 20000)

1-sample proportions test with continuity correction

data: 10231 out of 20000, null probability 0.5
X-squared = 10.626, df = 1, p-value = 0.001115
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.5045958 0.5184998
sample estimates:
      p
0.51155

> prop.test(x = 10014 + 10231, n = 40000)

1-sample proportions test with continuity correction

data: 10014 + 10231 out of 40000, null probability 0.5
X-squared = 5.978, df = 1, p-value = 0.01449
```

```

alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
  0.5012126 0.5110362
sample estimates:
  p
  0.506125

```

The pooled estimate of the proportion is  $\hat{p} = 50.6\%$ , rather close to the value predicted in Diaconis *et al.* (2007). In addition, the hypothesis  $H_0 : p = 0.5$  is rejected with significance level  $P = 0.01449$ .

What sample size is needed to obtain a margin of error of 0.5% using a 95% confidence interval? Methods for solving this type of problem will be discussed in Section 16.2. ■

## 15.6 Assumptions

The use of the normal distribution in the probability calculations is justified by the central limit theorem, although a large sample is needed for this purpose. To adapt the rule of thumb given previously for the normal approximation to the binomial, we will require that

$$n\hat{p} \geq 5$$

and

$$n(1 - \hat{p}) \geq 5$$

where  $n$  is the sample size.

For the two sample case we simply apply these conditions separately to the two samples.

# Chapter 16

## Sample Size Estimation for Confidence Intervals

Sampling design is a crucial part of any study involving the collection of data. The objectives of the study should be defined in advance, and the sampling scheme designed specifically to meet those objectives. Cost might be an important feature of the design. We may wish to avoid doing more sampling than is necessary to achieve the objective.

An important example of sampling design is the determination of an appropriate sample size for constructing a confidence interval. We next consider this problem for two kinds of confidence intervals previously studied.

### 16.1 General Approach to Sample Size Calculations: Normal Approximations

Recall that the confidence interval for a population mean, given population variance  $\sigma^2$  is

$$CI_{1-\alpha} = \bar{X}_n \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

Recall that the confidence interval for a population mean, given population variance  $\sigma^2$  is

$$CI_{1-\alpha} = \bar{X}_n \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

The margin of error is

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Now suppose before we collect the sample we decide that the margin of error should be  $E_o$ , and the confidence level should be  $(1 - \alpha)100\%$ . We can use

the previous expression to determine what the sample size should be, giving

$$n = \left( z_{\alpha/2} \frac{\sigma}{E_o} \right)^2$$

as the required sample size. As a technical note, the  $n$  calculated will not generally be an integer. In this case we would always round up, as opposed to rounding to the nearest integer. This way, we ensure that the confidence level reported is not overestimated.

Of course, we would rarely know the actual value  $\sigma^2$ , so we would have to substitute an estimate  $\hat{\sigma}^2$ , giving

$$n \approx \left( z_{\alpha/2} \frac{\hat{\sigma}}{E_o} \right)^2.$$

If we had a previous study, or some other prior knowledge, we could rely on that for  $\hat{\sigma}^2$ . Failing that, a reasonable alternative is to do an initial *pilot study*. This would be a small sample whose primary purpose would be to obtain an estimate of  $\hat{\sigma}^2$ . An estimate of the required sample size for a fixed margin of error could then be obtained, and the sample then completed.

**Example 16.1** A certain industrial process is designed to produce ball bearings of a certain diameter. It is decided to estimate the mean diameter of the process to within 0.01 centimeters with a confidence level of 99%. A pilot sample of 20 ball bearings has a sample standard deviation of 0.04 centimeters.

To calculate the total sample size needed to achieve the objectives we note that

$$\begin{aligned}\alpha &= 0.01 \\ z_{\alpha/2} &= 2.576 \\ \hat{\sigma} &= 0.04 \\ E_o &= 0.03\end{aligned}$$

so that the sample size required is

$$\begin{aligned}n &\approx \left( z_{\alpha/2} \frac{\hat{\sigma}}{E_o} \right)^2 \\ &= \left( 2.576 \frac{0.04}{0.01} \right)^2 \\ &= 106.17\end{aligned}$$

which, when rounded up, gives a sample size of 107 ball bearings. Note that 20 have already been collected, so we need an additional 87. ■

## 16.2 Sample Size for a Confidence Interval for a Population Proportion

The appropriate sample size required to estimate a population proportion can also be estimated using similar reasoning, except that there are some technical differences. Recall that the level  $(1 - \alpha)100\%$  confidence interval is given by

$$CI_{\alpha/2} = \hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

so that the margin of error is

$$E = z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}.$$

If this expression is rearranged we get

$$n = \hat{p}(1 - \hat{p}) \left( \frac{z_{\alpha/2}}{E} \right)^2.$$

Of course, a similar problem to that encountered for the population mean sample size calculation appears here, namely that we don't observe  $\hat{p}$  before the sample is collected. There are two possible solutions.

The first solution is to note that the largest possible value for  $p(1 - p)$  occurs when  $p = 1/2$ . Therefore, if we simply substitute  $1/2$  for  $\hat{p}$  in the previous expression we get an estimate for the sample size which is guaranteed to be large enough. Then for a given margin of error  $E_o$  and confidence level  $(1 - \alpha)100\%$  an estimate for the required sample size would be

$$\begin{aligned} n &= 1/2(1 - 1/2) \left( \frac{z_{\alpha/2}}{E_o} \right)^2 \\ &= \left( \frac{z_{\alpha/2}}{2E_o} \right)^2 \end{aligned}$$

This procedure would be appropriate if it is expected that the true population proportion  $p$  is not too far from  $1/2$ . This is usually the case with opinion surveys.

The second solution would be to somehow obtain an initial estimate  $p^*$  of the population proportion  $p$ . Then an estimate of the required sample size would be

$$n = p^*(1 - p^*) \left( \frac{z_{\alpha/2}}{E_o} \right)^2.$$

The estimate  $p^*$  could be estimated from an initial pilot study, or from some other prior knowledge. If it is believed that  $p$  is relatively small, then  $p^*$  should be set to be the largest value that  $p$  could feasibly take.

**Example 16.2** Many opinion surveys report a margin of error of 3% with a 95% confidence level. In this case we have

$$\begin{aligned} E_o &= 0.03 \\ z_{\alpha/2} &= 1.96 \end{aligned}$$

We usually expect proportions in opinion surveys to be close to 1/2, so to obtain the necessary sample size we use the equation

$$\begin{aligned} n &= \left( \frac{z_{\alpha/2}}{2E_o} \right)^2 \\ &= \left( \frac{1.96}{2 \times 0.03} \right)^2 \\ &= 1067.1. \end{aligned}$$

Rounding up to the next highest integer gives a required sample size of 1068. This is a typical sample size of many published opinion surveys. ■

### 16.3 Sample Size for a Confidence Interval for Differences in Means

Recall that the confidence interval for a difference in population means, given population variances  $\sigma_1^2$  and  $\sigma_2^2$ , and sample sizes  $n_1$  and  $n_2$  is

$$CI_{1-\alpha} = \bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

The margin of error is therefore

$$E = z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

We can, in principle, predict  $E$  for any configuration  $(\sigma_1^2, \sigma_2^2, n_1, n_2)$ .

It is somewhat more difficult to invert the calculation, to obtain the sample sizes needed for a fixed margin of error  $E$ . We may always specify that  $n = n_1 = n_2$ , which is referred to as a *balanced design*. In this case we can obtain the formula

$$n \approx \left( \frac{z_{\alpha/2}}{E} \right)^2 (\sigma_1^2 + \sigma_2^2).$$

**Example 16.3** Suppose we are given  $\sigma_1^2 = 3.4$ ,  $\sigma_2^2 = 5.7$ , and we wish to determine a sample size  $n$  (per sample) which will yield a margin of error of 1.2 for an estimate of a difference in means ( $\alpha = 0.05$ ). We get directly,

$$\begin{aligned} n &\approx \left( \frac{z_{\alpha/2}}{E} \right)^2 (\sigma_1^2 + \sigma_2^2) \\ &= \left( \frac{1.96}{1.2} \right)^2 (3.4 + 5.7) \\ &= 24.3. \end{aligned}$$

So we use, conservatively, a sample size of 25 per sample. ■

**Example 16.4** Suppose in the previous example, accept a total sample size of 50, but decide to vary the allocation. In the balanced design for  $n_1 = 25$ ,  $n_2 = 25$  we have

$$\begin{aligned} E &= z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \\ &= 1.96 \sqrt{\frac{3.4}{25} + \frac{5.7}{25}} \\ &= 1.18, \end{aligned}$$

slightly less than the target margin of error of the previous example. The following sample sizes yield the following margins of error

$$\begin{aligned} (n_1, n_2) &= (40, 10) \Rightarrow E = 1.586 \\ (n_1, n_2) &= (30, 20) \Rightarrow E = 1.237 \\ (n_1, n_2) &= (20, 30) \Rightarrow E = 1.176 \\ (n_1, n_2) &= (10, 40) \Rightarrow E = 1.361. \end{aligned}$$

Notice that the balanced design is *not* optimal. This is because the variances are unequal. In this case, the inference can be improved by allocating more sample size to the population with the higher variance. It can be shown that the optimal allocation, if achievable, satisfies  $\sigma_1/\sigma_2 = n_1/n_2$ . ■

# Chapter 17

## Power curves

We introduced earlier the notion of *power*. For example, given hypotheses

$$\begin{aligned} H_o : \mu &\leq \mu_0 \\ H_a : \mu &> \mu_0 \end{aligned}$$

a Type II error occurs when  $H_o$  is not rejected when  $H_a$  is true. Recall that  $H_a$  is a composite. If we wish to calculate the probability of a Type II error, we must specify exactly which value of  $\mu > \mu_0$  is the correct one. Usually, the notation

$$\begin{aligned} \beta(\mu) &= P(\text{ Type II error }) \\ &= P(Z_{obs} < z_\alpha | \mu) \end{aligned}$$

is used. We then define

$$\begin{aligned} \text{Power}(\mu) &= 1 - \beta(\mu) \\ &= P(\text{ correctly rejecting } H_o \text{ when the true mean is } \mu) \end{aligned}$$

again noting that the exact value of the power depends on the exact value of  $\mu$ .

We have taken the Type I error to be a single number associated with  $H_o$ , but have allowed the Type II error to vary over  $H_a$ . It is therefore natural to investigate the resulting functional form. We have defined  $\alpha(\theta) = P(\tilde{X} \in R | \theta)$ , and from it derive both the Type I and Type II errors. Suppose we consider the general upper tailed test

$$\begin{aligned} H_o : \mu &\leq \mu_0 \\ H_a : \mu &> \mu_0 \end{aligned}$$

based on an *iid* normal sample from  $N(\mu, \sigma^2)$  of size  $n$ . Then we use statistic

$$Z_{obs} = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}},$$

and reject  $H_o$  with significance level  $\alpha$  with rejection region

$$R = \{Z_{obs} \geq z_\alpha\}.$$

Then

$$\begin{aligned}\alpha(\mu) &= P(Z_{obs} \geq z_\alpha \mid \mu) \\ &= P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} + \sqrt{n}\frac{(\mu - \mu_0)}{\sigma} \geq z_\alpha \mid \mu\right) \\ &= P\left(Z \geq z_\alpha - \sqrt{n}\frac{(\mu - \mu_0)}{\sigma}\right)\end{aligned}\tag{17.1}$$

where  $Z \sim N(0, 1)$ . A few things are worth noting about (17.1). First, a plot of  $\alpha(\mu)$  is given in Figure 17.1 for parameters  $\mu_0 = 100$ ,  $\sigma = 10$ ,  $n = 50$ . Note that  $\alpha(100) = 0.05$  as expected, and that this is the maximum value of  $\alpha(\mu)$  within  $H_o$ , and is therefore the appropriate choice for Type I error. Next, note that  $\alpha(\mu)$  increases above  $\mu = \mu_0 = 100$ ,  $\alpha(\mu)$  also increases, and does so quite sharply just before  $\mu = 105$ , reaching a value very close to 1 for all values  $\mu > 105$ . Clearly, the power of the test should depend on the distance of a particular alternative hypothesis from the null hypothesis  $\mu_0$ .

Letting  $\Phi$  be the CDF of the standard normal distribution  $N(0, 1)$  we can write:

$$\alpha(\mu) = 1 - \Phi\left(z_\alpha - \sqrt{n}\frac{(\mu - \mu_0)}{\sigma}\right).\tag{17.2}$$

Since  $\Phi(x)$  is an increasing function of  $x$ , for  $\mu > \mu_0$ , we also have that  $\alpha(\mu)$  has an increasing relationship with

$$\sqrt{n}\delta, \text{ where } \delta = \frac{(\mu - \mu_0)}{\sigma}.$$

That power should increase with sample size  $n$  is expected. Also expected is that power increases with the quantity  $\mu - \mu_0$ . However, from the point of view of power, it is the ratio  $\delta = (\mu - \mu_0)/\sigma$  which is important. This is can be interpreted as the size of the effect of interest in units of standard deviation.

## 17.1 Power Analysis and the Noncentral $t$ -distribution

There is clearly an advantage to knowing  $\sigma$  when estimating sample size. However, in practice  $\sigma$  is usually unknown. While it may be possible to substitute an estimate, and even to bound the error of this estimate this doesn't completely address the problem when small sample sizes are being planned, and therefore a  $t$ -test is being proposed.

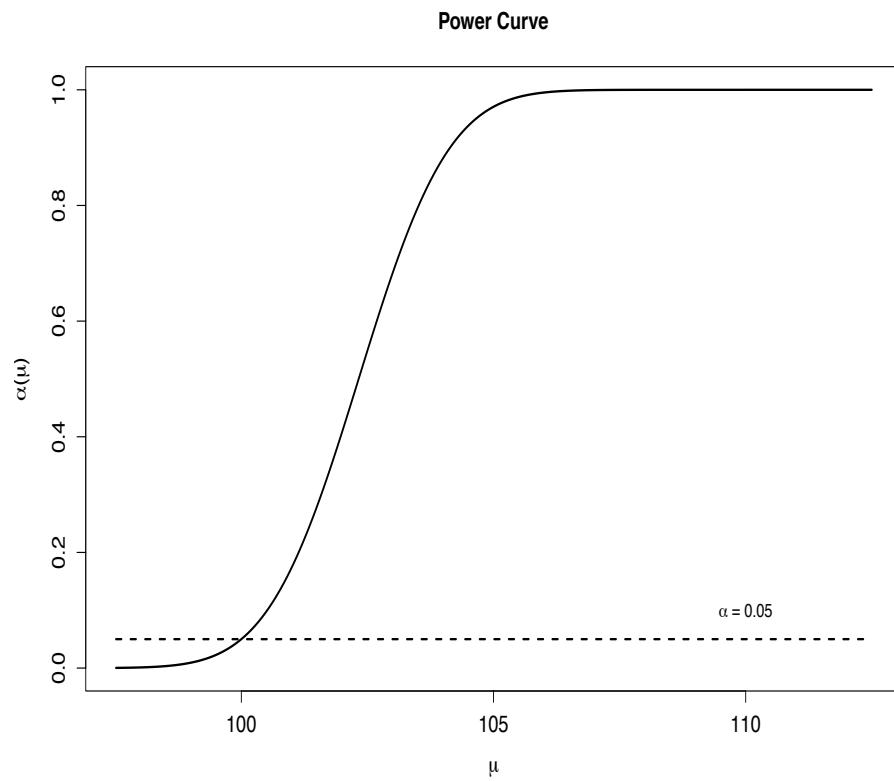


Figure 17.1: Example of  $\alpha(\mu)$  for parameters  $\mu_0 = 100$ ,  $\sigma = 10$ ,  $n = 50$ .

Recall from Section 4.15 that the  $t$ -distribution with  $\nu$  degrees of freedom is equivalent to

$$T = \frac{Z}{\sqrt{W/\nu}} \quad (17.3)$$

where  $Z \sim N(0, 1)$ ,  $W \sim \chi_{\nu}^2$  and  $Z \perp W$ .

The *noncentral t*-distribution is obtained by setting

$$T_{\eta} = \frac{Z + \eta}{\sqrt{W/\nu}}$$

for a constant  $\eta$  (or equivalently, by replacing  $Z$  in the numerator of (17.3) with  $Z_{\eta} \sim N(\eta, 1)$ ). The original  $t$ -distribution ( $\eta = 0$ ) may then be referred to as the *central t*-distribution. The value  $\eta$  is the *noncentrality parameter*.

The importance of  $T_{\eta}$  is that it models the distribution of a test statistic under an alternative hypothesis. Suppose we are given null mean  $\mu_0$  and alternative  $\mu$ . The  $T$ -statistic is

$$\begin{aligned} T &= \sqrt{n} \frac{\bar{X} - \mu_0}{S} \\ &= \frac{\sqrt{n}(\bar{X} - \mu)/\sigma + \sqrt{n}(\mu - \mu_0)/\sigma}{S/\sigma}. \end{aligned}$$

If  $\mu$  is the true mean we have

$$T = \frac{Z + \sqrt{n}\delta}{\sqrt{W/(n-1)}} \text{ where } \delta = (\mu - \mu_0)/\sigma,$$

where  $Z \sim N(0, 1)$ ,  $W \sim \chi_{n-1}^2$  and  $Z \perp W$ , so that  $T$  possesses a noncentral  $t$ -distribution with  $n-1$  degrees of freedom and noncentrality parameter  $\eta = \sqrt{n}\delta$ . For an upper tailed test, a level  $\alpha$  rejection region is  $T \geq t_{n-1,\alpha}$ , therefore the probability of a type II error for alternative  $\mu$  is

$$\beta(\delta) = P(T_{\sqrt{n}\delta} \leq t_{n-1,\alpha}) \text{ where } \delta = (\mu - \mu_0)/\sigma.$$

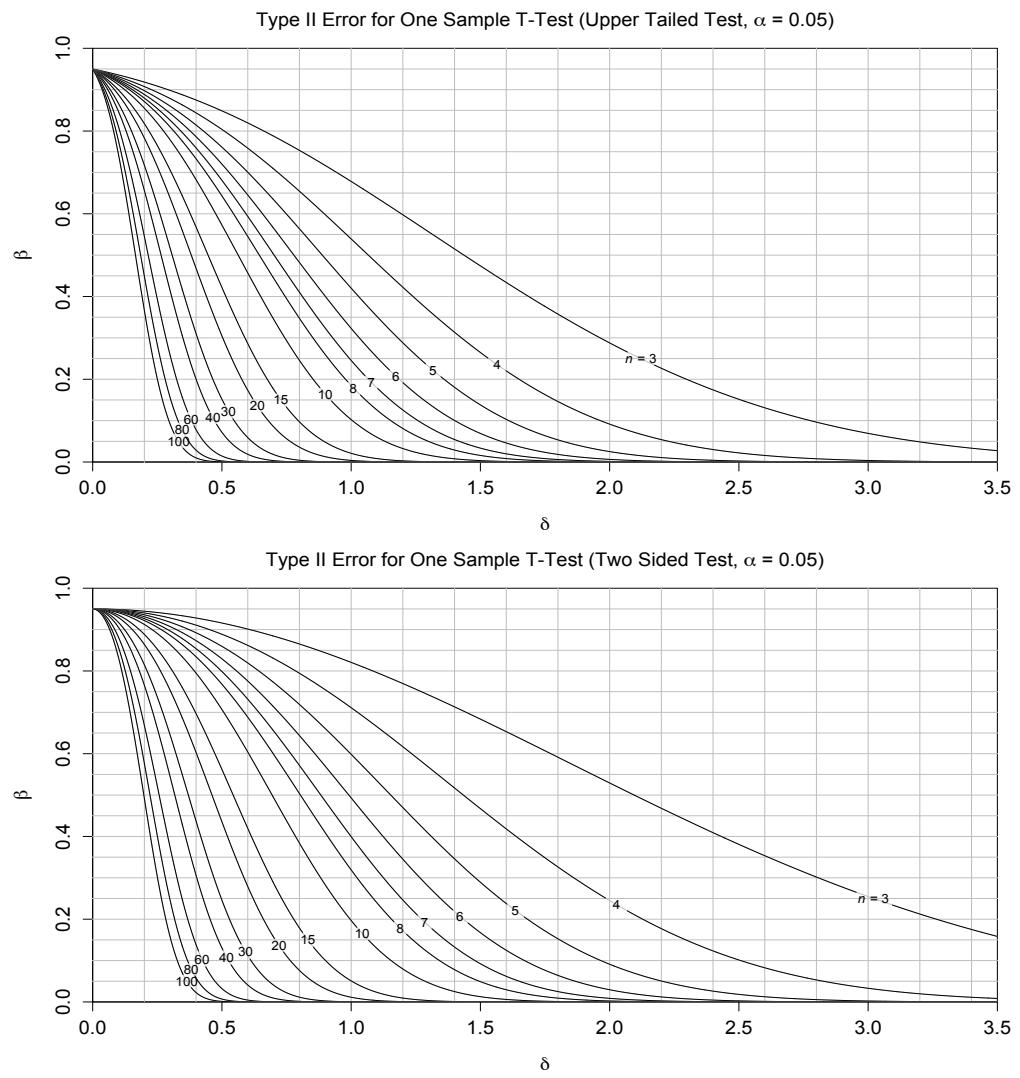
The CDF of  $T_{\eta}$  is available in most statistical computing environments. For a two sided test we have

$$\beta(\delta) = P(T_{\sqrt{n}\delta} \leq t_{n-1,\alpha/2}) - P(T_{\sqrt{n}\delta} \leq -t_{n-1,\alpha/2}) \text{ where } \delta = (\mu - \mu_0)/\sigma.$$

## 17.2 Sample Proportion

The test statistic is

$$Z_{obs} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

Figure 17.2: Power analysis curves for one sample  $t$ -test.

Given alternative  $p$  we may write rewrite  $Z_{obs}$  as, equivalently

$$\begin{aligned} Z_{obs} &= \frac{\hat{p} - p + p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \\ &= \frac{Z \sqrt{\frac{p(1-p)}{n}} + p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \end{aligned}$$

where (approximately)  $Z \sim N(0, 1)$ . For an upper tailed test, we reject if  $Z_{obs} \geq z_\alpha$ , so the probability of a type II error becomes

$$\beta_U(p, p_0, n, \alpha) = \Phi \left( \frac{z_\alpha \sqrt{p_0(1-p_0)} - \sqrt{n}(p - p_0)}{\sqrt{p(1-p)}} \right). \quad (17.4)$$

An equation such as (17.4) allows the estimation of a type II error as a function of sample size for a suitable alternative hypothesis. The relation can be inverted to give the sample size required for a fixed power. To do this, note the equality

$$\beta = \Phi(-z_\beta),$$

and compare to (17.4). This yields the equality

$$-z_\beta = \frac{z_\alpha \sqrt{p_0(1-p_0)} - \sqrt{n}(p - p_0)}{\sqrt{p(1-p)}},$$

which may be rewritten as

$$n(p, p_0, \alpha, \beta) = \left( \frac{z_\alpha \sqrt{p_0(1-p_0)} + z_\beta \sqrt{p(1-p)}}{(p - p_0)} \right)^2. \quad (17.5)$$

The lower tailed test has rejection region  $Z_{obs} \leq -z_\alpha$  so a similar argument gives type II error

$$\beta_L(p, p_0, n, \alpha) = 1 - \Phi \left( \frac{-z_\alpha \sqrt{p_0(1-p_0)} - \sqrt{n}(p - p_0)}{\sqrt{p(1-p)}} \right), \quad (17.6)$$

which also yields the same equation (17.5) for estimated sample size as for the upper tailed test.

Finally, for a two sided test, the rejection region is  $|Z_{obs}| \geq z_{\alpha/2}$  which, by a similar argument, yields

$$\begin{aligned} \beta_{two}(p, p_0, n, \alpha) &= \Phi \left( \frac{z_{\alpha/2} \sqrt{p_0(1-p_0)} - \sqrt{n}(p - p_0)}{\sqrt{p(1-p)}} \right) \\ &\quad - \Phi \left( \frac{-z_{\alpha/2} \sqrt{p_0(1-p_0)} - \sqrt{n}(p - p_0)}{\sqrt{p(1-p)}} \right) \\ &= \beta_U(p, p_0, n, \alpha/2) + \beta_L(p, p_0, n, \alpha/2) - 1. \quad (17.7) \end{aligned}$$

Inversion of (17.7) is not as straightforward as for the upper and lower tailed tests, but can be accomplished numerically using a suitable statistical computing tool. In addition, a very good approximation is available. The two sided rejection region is the union of disjoint events  $R_L = \{Z_{obs} \leq -z_{\alpha/2}\}$  and  $R_U = \{Z_{obs} \geq z_{\alpha/2}\}$ . Suppose we have alternative  $p > p_0$ , and a reasonably large sample size  $n$ . We would then have  $P(R_L) \approx 0$ . But

$$\beta_L(p | p_0, n, \alpha/2) = P(R_L^c) = 1 - P(R_L),$$

which means

$$\beta_{two}(p, p_0, n, \alpha) \approx \beta_U(p | p_0, n, \alpha/2) \text{ for } p > p_0 \quad (17.8)$$

and similarly

$$\beta_{two}(p, p_0, n, \alpha) \approx \beta_L(p | p_0, n, \alpha/2) \text{ for } p < p_0, \quad (17.9)$$

so that a sample size estimate can be obtained from  $n(p, p_0, \alpha/2, \beta)$ . In essence, the power analysis for two sided test of size  $\alpha$  can be reasonably undertaken by regarding it as the appropriate one sided test of size  $\alpha/2$ .

# Chapter 18

## Inference for Variances

Typically, inference is concerned with population means or proportions, or with differences in these quantities. We have seen that population variance plays a crucial role in this type of inference, but is usually not the object of the inference. Of course, sometimes it will be important to estimate a population variance. In addition there will often be interest in establishing whether or not two population variances are equal.

### 18.1 Inference for a Single Variance

Recall the discussion of the sample variance in Section 12.3. Suppose  $X_1, \dots, X_n$  is an *iid* sample from normal distribution  $N(\mu, \sigma^2)$ . The sample variance is:

$$S_n^2 = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{n-1}.$$

If we set

$$W = \frac{(n-1)S_n^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\sigma^2},$$

then it may be shown that

$$W \sim \chi_{n-1}^2, \quad (18.1)$$

that is,  $W$  has a  $\chi^2$  distribution with  $n-1$  degrees of freedom (see Section 4.15).

Next, we can define critical values  $\chi_{\nu,\alpha}$  (obtainable from Table A.4) which satisfy

$$P(W > \chi_{\nu,\alpha}^2) = \alpha,$$

when  $W \sim \chi_{\nu}^2$ . Figure 18.1 shows critical values  $\chi_{10,0.975}^2 \approx 3.25$  and  $\chi_{10,0.025}^2 \approx 20.48$ .

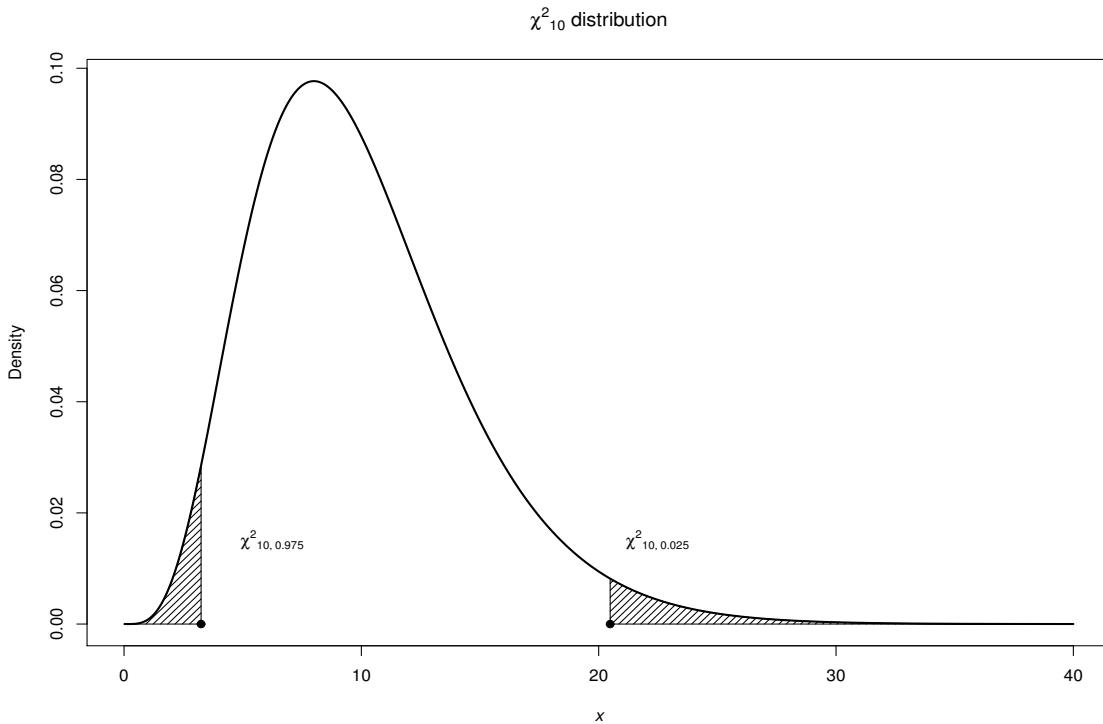


Figure 18.1: Critical values  $\chi^2_{10,0.975} \approx 3.25$  and  $\chi^2_{10,0.025} \approx 20.48$  for a  $\chi^2$  distribution with 10 df. The shaded areas each represent a probability of 0.025.

We can use this distribution to construct a level  $(1 - \alpha)$  confidence interval for a variance of a normally distributed population. We first write

$$P\left(\chi^2_{n-1,1-\alpha/2} < \frac{(n-1)S_n^2}{\sigma^2} < \chi^2_{n-1,\alpha/2}\right) = 1 - \alpha. \quad (18.2)$$

The inequalities may be rewritten

$$\frac{S_n^2}{\left(\chi^2_{n-1,\alpha/2}\right)/(n-1)} < \sigma^2 < \frac{S_n^2}{\left(\chi^2_{n-1,1-\alpha/2}\right)/(n-1)}, \quad (18.3)$$

which, by (18.2), becomes a level  $(1 - \alpha)$  confidence interval for  $\sigma^2$ .

**Example 18.1** A sample of size  $n = 10$  is sampled from a  $N(\mu = 10, \sigma^2 = 10.24)$  distribution:

18.42 19.95 21.42 22.57 23.35 25.98 26.31 26.86 27.38 31.35

(This data was generated by a random number generator). To construct a 95% confidence interval we need the following quantities:

$$\begin{aligned} S^2 &= 15.36, \\ \chi_{n-1,1-\alpha/2}^2 &= \chi_{9,0.975}^2 \approx 2.70, \\ \chi_{n-1,\alpha/2}^2 &= \chi_{9,0.025}^2 \approx 19.02, \end{aligned}$$

where the critical values are obtained from Table A.4. Substituting into (18.3) yields

$$\frac{15.36}{19.02/9} < \sigma^2 < \frac{15.36}{2.70/9}, \quad (18.4)$$

or

$$7.27 < \sigma^2 < 51.20, \quad (18.5)$$

which contains the true variance  $\sigma^2 = 10.24$ . ■

The confidence interval for  $\sigma^2$  can often seem large. Of course, it sometimes will be, but the fact that  $\sigma^2$  is in units squared can give a misleading picture. Because all quantities in (18.3) are positive, we can take the square root of each quantity, with the confidence interval (now for the standard deviation) remaining valid:

$$\frac{S_n}{\sqrt{(\chi_{n-1,\alpha/2}^2)/(n-1)}} < \sigma < \frac{S_n}{\sqrt{(\chi_{n-1,1-\alpha/2}^2)/(n-1)}}, \quad (18.6)$$

except that now all quantities are in the units of interest. The 95% confidence interval (18.7) of Example 18.1 can be equivalently written:

$$2.70 < \sigma < 7.16, \quad (18.7)$$

which contains the true standard deviation  $\sigma = \sqrt{10.24} = 3.2$ .

## 18.2 Upper Confidence Bounds

Suppose we are using a pilot study to estimate a variance  $\sigma^2$  to estimate sample sizes for a larger study. In this case, we may be more interested in saying

$$\sigma \text{ is no larger than } \sigma_U$$

instead of

$$\sigma \text{ is between } \sigma_L \text{ and } \sigma_U.$$

Why would this difference be important? Suppose instead of

$$P\left(\chi_{n-1,1-\alpha/2}^2 < \frac{(n-1)S_n^2}{\sigma^2} < \chi_{n-1,\alpha/2}^2\right) = 1 - \alpha, \quad (18.8)$$

we wrote

$$P\left(\chi_{n-1,1-\alpha}^2 < \frac{(n-1)S_n^2}{\sigma^2}\right) = 1 - \alpha. \quad (18.9)$$

Both statements are true, but the second allows us to assign a confidence level of  $(1 - \alpha)$  to the statement

$$\sigma^2 < \frac{S_n^2}{\left(\chi_{n-1,1-\alpha}^2\right)/(n-1)} \quad (18.10)$$

or

$$\sigma < \frac{S_n}{\sqrt{\left(\chi_{n-1,1-\alpha}^2\right)/(n-1)}}. \quad (18.11)$$

We call (18.11) an *upper confidence bound* for  $\sigma$ . The advantage is that  $\chi_{n-1,1-\alpha}^2 > \chi_{n-1,1-\alpha/2}^2$ , so that a level  $(1 - \alpha)$  upper confidence bound is always smaller than the upper bound of a level  $(1 - \alpha)$  confidence interval.

**Example 18.2** From Example 18.1 we had a 95% confidence interval

$$2.70 < \sigma < 7.16. \quad (18.12)$$

To construct a 95% upper confidence bound for  $\sigma$ , we need

$$\chi_{n-1,1-\alpha}^2 = \chi_{9,0.95}^2 \approx 3.33,$$

which, when substituted into (18.11) gives

$$\sigma < \frac{S_n}{\sqrt{\left(\chi_{n-1,1-\alpha}^2\right)/(n-1)}} \approx \frac{3.92}{\sqrt{3.33/9}} = 6.45. \quad (18.13)$$

The upper confidence bound 6.45 compares to 7.16, the upper bound of a confidence interval of the same level. ■

### 18.3 Sample Size Estimation

We calculate confidence intervals and confidence bounds with the expression,

$$S_n \times \frac{1}{\sqrt{\left(\chi_{n-1,p}^2\right)/(n-1)}}$$

where  $p = \alpha/2, 1 - \alpha, \alpha/2$ , as needed, so it's worth examining the normalized critical values:

$$\sqrt{\frac{\chi_{\nu,p}^2}{\nu}},$$

shown in Figure 18.2. Clearly, we wish

$$\sqrt{\frac{\chi_{\nu,p}^2}{\nu}} \approx 1.$$

The sample size is given by  $\nu = n - 1$ . Clearly, once sample sizes are above, say,  $n = 50$ ,  $S_n$  is reasonable close to  $\sigma$ , in particular, we have a 95% confidence interval approximately

$$S_n/1.2 < \sigma < S_n/0.8.$$

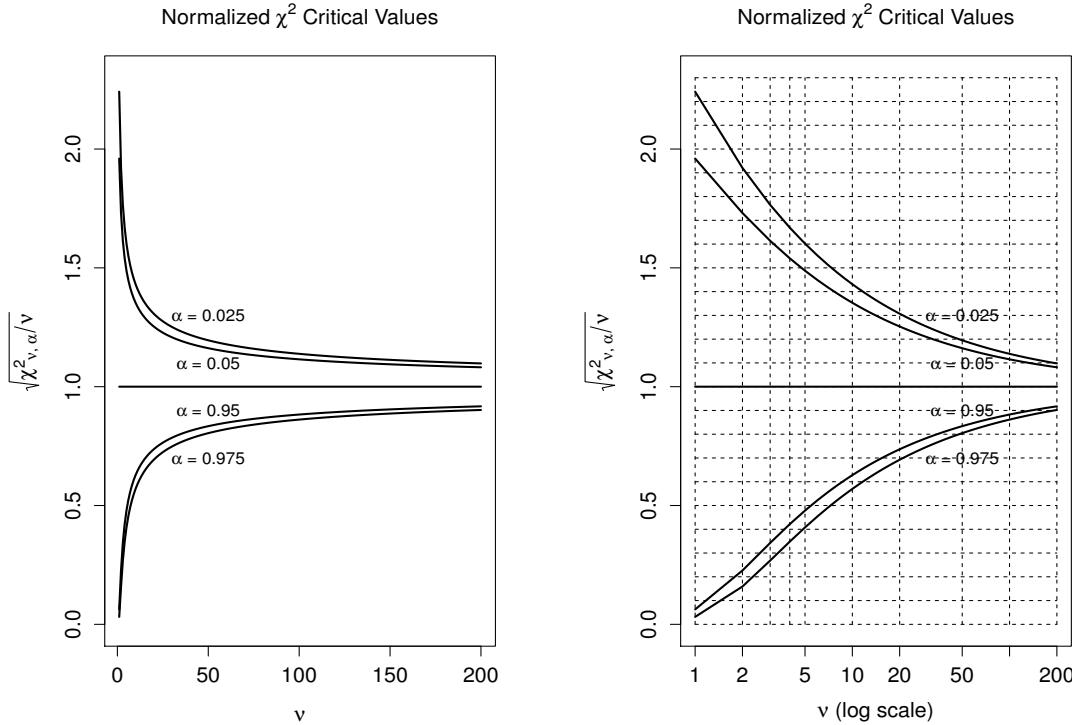


Figure 18.2: Normalized critical values  $\sqrt{\chi_{\nu,\alpha}^2}$ .

## 18.4 IQR as Estimate of Variance

Recall that the IQR (Section 9.3) is defined by

$$IQR = Q_3 - Q_1 = Q(75\%) - Q(25\%).$$

For a standard normal distribution we have

$$Q(75\%) = -Q(25\%) = z_{0.25} = 0.6745,$$

so that

$$IQR = Q_3 - Q_1 = 1.349.$$

Recall the rule for general normal quantiles,

$$X_p = \mu + \sigma Z_p$$

where  $Z_p$  and  $X_p$  are the  $p$ th quantiles for a  $N(0, 1)$  and  $N(\mu, \sigma^2)$  distribution, respectively. Therefore, for a general normal distribution  $N(\mu, \sigma^2)$  we have

$$IQR = 1.349 \times \sigma \text{ or } \sigma = IQR/1.349,$$

which can serve as an alternative estimate of  $\sigma$ .

**Example 18.3** From Example 18.1 we have (from the medians of the upper and lower halves of the data)

$$\begin{aligned} Q(25\%) &= X_{(3)} = 21.42, \\ Q(75\%) &= X_{(8)} = 26.86, \\ IQR &= 26.86 - 21.42 = 5.4. \end{aligned}$$

This gives estimate

$$\sigma \approx IQR/1.349 = 5.4/1.349 = 4.00,$$

which is quite close to the sample standard deviation  $S = 3.92$ . ■

## 18.5 Hypothesis Tests for Two Population Variances

We return to the two population structure,

	Pop'n 1	Pop'n 2
Population mean	$\mu_1$	$\mu_2$
Population variance	$\sigma_1^2$	$\sigma_2^2$
Sample size	$n_1$	$n_2$
Sample mean	$\bar{X}_1$	$\bar{X}_2$
Sample variance	$S_1^2$	$S_2^2$

Except that now we are concerned with hypotheses of the form

1. One sided, lower tailed hypothesis

$$\begin{aligned} H_o : \sigma_1^2 &\geq \sigma_2^2 \\ H_a : \sigma_1^2 &< \sigma_2^2 \end{aligned}$$

We are looking for evidence that  $\sigma_1^2$  is **less than**  $\sigma_2^2$ .

2. One sided, upper tailed hypothesis

$$\begin{aligned} H_o : \sigma_1^2 &\leq \sigma_2^2 \\ H_a : \sigma_1^2 &> \sigma_2^2 \end{aligned}$$

We are looking for evidence that  $\sigma_1^2$  is **greater than**  $\sigma_2^2$ .

3. Two sided hypothesis

$$\begin{aligned} H_o : \sigma_2^2 &= \sigma_1^2 \\ H_a : \sigma_2^2 &\neq \sigma_1^2 \end{aligned}$$

We are looking for evidence that  $\sigma_1^2$  is **not equal to**  $\sigma_2^2$ .

A hypothesis test can be based on the fact that the statistic

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1, n_2-1}$$

has an  $F$ -distribution with  $n_1 - 1$  numerator degrees of freedom and  $n_2 - 1$  denominator degrees of freedom, following the discussion in Section 4.15 and noting the equality in (18.1). If  $\sigma_1^2 = \sigma_2^2$ , then

$$F = \frac{S_1^2}{S_2^2} \sim F_{n_1-1, n_2-1}.$$

Given this test statistic, the observed significance levels are calculated according to the following rules.

1. One sided, lower tailed hypothesis

$$\alpha_{obs} = P(F < F_{obs})$$

2. One sided, upper tailed hypothesis

$$\alpha_{obs} = P(F > F_{obs})$$

3. Two sided hypothesis

$$\alpha_{obs} = 2 \min(P(F < F_{obs}), P(F > F_{obs})).$$

where we assume  $F \sim F_{n_1-1, n_2-1}$ , and set

$$F_{obs} = \frac{S_1^2}{S_2^2} \sim F_{n_1-1, n_2-1}.$$

There will be three different forms of the rejection region depending on the form of the hypothesis.

1. One sided, lower tailed hypothesis

$$\text{Reject } H_o \text{ if } F_{obs} \leq F_{n_1-1, n_2-1, 1-\alpha}$$

2. One sided, upper tailed hypothesis

$$\text{Reject } H_o \text{ if } F_{obs} \geq F_{n_1-1, n_2-1, \alpha}$$

3. Two sided hypothesis

$$\text{Reject } H_o \text{ if } F_{obs} \leq F_{n_1-1, n_2-1, 1-\alpha/2} \text{ or } F_{obs} \geq F_{n_1-1, n_2-1, \alpha/2}$$

where  $F_{n_1-1, n_2-1, \alpha}$  is the  $\alpha$  critical value from an  $F_{n_1-1, n_2-1}$  distribution. These values can be obtained from Tables A.5-A.8.

The  $F$ -distribution possesses the following symmetry:

$$F_{\nu_1, \nu_2, 1-\alpha} = 1/F_{\nu_2, \nu_1, \alpha},$$

so that many tables include only the upper tail critical values.

**Example 18.4** We are given samples of size  $n_1 = 10$  and  $n_2 = 20$  from populations with distributions  $N(\mu_1 = 20, \sigma_1^2 = 1.44)$  and  $N(\mu_2 = 30, \sigma_2^2 = 9)$ . Suppose we observe sample variances

$$\begin{aligned} S_1^2 &= 1.33, \\ S_2^2 &= 9.97, \\ F_{obs} &= 1.33/9.97 = 0.1334. \end{aligned}$$

The relevant critical values (Tables A.5-A.8) for a two-sided size  $\alpha = 0.05$  test are

$$\begin{aligned} F_{n_1-1, n_2-1, 1-\alpha/2} &= F_{9, 19, 0.975} \approx 0.2715 \\ F_{n_1-1, n_2-1, \alpha/2} &= F_{9, 19, 0.025} \approx 2.880. \end{aligned}$$

Since  $F_{obs} < F_{9, 19, 0.975}$  we reject the hypothesis  $\sigma_1^2 = \sigma_2^2$ . ■

An R function for *Bartlett's test*, a general test for equality of variances will be discussed in Section 24.1.

## 18.6 Assumptions

The data are assumed to be normally distributed. For samples about  $n \geq 30$  the procedures will work well if there is minimal skewness. In general, we should expect  $IQR/1.349 \approx S_n$ .

# Chapter 19

## Nonparametric Inference

Many statistical procedures make very strict assumptions about the distribution of a sample, specifying its exact form (normal, binomial, Poisson and so on). Once the form of the distribution is given, inference can be focused on the *parameters* ( $\mu, \sigma^2$  for a normal distribution,  $p$  for a binomial distribution, and so on). The use of the  $t$ -distribution,  $\chi^2$  distribution and  $F$ -distribution are usually based on assumptions of normality. The correctness of these procedures depends on the validity of the assumptions. A procedure may be *robust* to assumptions, in the sense that it may still be approximately correct when the assumptions are not strictly met. Even then, we must still make sure that the assumption violations are not too severe. For example, a procedure which assumes normality may still be accurate when the sample distribution is symmetric but not normal, but not when the sample distribution is severely skewed.

It is possible to devise statistical procedures that do not require such strict assumptions about the form of a density. These are referred to as *nonparametric* statistical procedures (the term *distribution-free* is sometimes used in this context). Although they still rely on assumptions such as independence (usually) or distributional symmetry (sometimes), they do not require a specific density form, and are exactly correct as long as these more minimal assumptions are met.

### 19.1 Sign Test

This example is from Maskin *et al.*, 1985 in the journal *Circulation*. The following table shows the measurement of heart rate (in beats per minute) before and 30 minutes after administering enalaprilat to 9 subjects.

Subject	Time after treatment 0 mins	Time after treatment 30 mins	Change in heart rate	+/-
1	96	92	-4	-
2	110	106	-4	-
3	89	86	-3	-
4	95	78	-17	-
5	128	124	-4	-
6	100	98	-2	-
7	72	68	-4	-
8	79	75	-4	-
9	100	106	+6	+

We are interested in determining whether or not the treatment has an effect on heart rate (that is, heart rate increases or decreases). Earlier, we treated this as paired sample, and examining the 9 differences

$$D_i = X_i - Y_i$$

found that we could reject a one sided hypothesis  $H_o : d \geq 0$  with observed significance level  $0.025 < \alpha_{obs} < 0.05$ . This means we would fail to reject a two sided hypothesis  $H_o : d = 0$  with  $\alpha_{obs} \leq 0.05$ .

Next, consider the fact that of the 9 subjects, 8 experienced a decrease in heart rate. If we let  $T$  equal the number of increases (or plus signs), then  $T = 1$ . If there is no effect of the treatment on heart rate, we might expect a subject's heart rate to be equally likely to increase or decrease, in which case  $T \sim bin(n, p)$  where  $n = 9$ ,  $p = 1/2$  (although even this assumption would need to be considered carefully). If we define null hypothesis

$$H_o : p = 1/2$$

and consider the two-sided hypothesis  $H_a : p \neq 1/2$ , then evidence at least a contradictory of  $H_o$  would take form

$$T \leq 1 \text{ or } T \geq 8$$

which, under the null hypothesis, has probability

$$\begin{aligned} P(T \leq 1 \text{ or } T \geq 8) &= P(T \leq 1) + P(T \geq 8) \\ &= P(T = 0) + P(T = 1) + P(T = 8) + P(T = 9) \\ &= \binom{9}{0}(1 - 1/2)^9 + \binom{9}{1}(1/2)^1(1 - 1/2)^8 \\ &\quad + \binom{9}{8}(1/2)^8(1 - 1/2)^1 + \binom{9}{9}(1/2)^9 \\ &= (1 + 9 + 9 + 1) \times (1/2)^9 \\ &= 20/2^9 \\ &\approx 0.039 \end{aligned}$$

therefore we have  $P$ -value  $\alpha_{obs} = 0.039$ . If we test against one-sided alternative  $H_a : p < 1/2$  we calculate  $\alpha_{obs}$  using

$$\begin{aligned} P(T \leq 1) &= P(T \leq 1) \\ &= P(T = 0) + P(T = 1) \\ &= \binom{9}{0}(1 - 1/2)^9 + \binom{9}{1}(1/2)^1(1 - 1/2)^8 \\ &= (1 + 9) \times (1/2)^9 \\ &= 10/2^9 \\ &\approx 0.0195, \end{aligned}$$

giving  $P$ -value  $\alpha_{obs} = 0.0195$ . We obtain a smaller  $P$ -value using a sign test than was obtained using the  $t$ -distribution.

The sign test is usually stated in terms of the median of  $D$ . Recall from the definition of *median* that

$$H_o : p = 1/2 \text{ is equivalent to } H_o : \text{median of } D = 0.$$

## 19.2 Signed Rank Test

The sign test is widely applicable, but is often not as powerful as other procedures. One strategy is to replace observations with suitably defined ranks. Suppose we define the null hypothesis for paired differences  $D$

$$H_o : \text{median of } D = 0.$$

which we test using the following paired difference data:

$D_i$	$ D_i $	Rank	$+/-$
-3.4	3.4	3	-
-1.4	1.4	1	-
3.2	3.2	2	+
4.3	4.3	4	+
5.9	5.9	5	+

For each value we calculate  $|D_i|$ , and then rank these values. As in the sign test, we assign ‘+’ or ‘-’ according to whether the value is above or below the hypothetical media. We can designate each rank as *positive* or *negative* according to its sign. We then set

$$\begin{aligned} T_+ &= \text{sum of positive ranks,} \\ T_- &= \text{sum of negative ranks.} \end{aligned}$$

In our example we have

$$\begin{aligned} T_+ &= 2 + 4 + 5 = 11, \\ T_- &= 1 + 3 = 4. \end{aligned}$$

Notice that we always have, for  $n$  data:

$$T_- + T_+ = 1 + 2 + \dots + n = n(n + 1)/2.$$

If  $H_o$  is true, and the distribution of  $X$  is distributed symmetrically about  $\tilde{\mu}_0$ , we would expect  $E[T_-] = E[T_+] = n(n + 1)/4$ . On the other hand, if median of  $D > 0$  we would expect  $T_+ > T_-$ , and if median of  $D < 0$  we would expect  $T_+ < T_-$ .

Let

$$T_{obs} = \min\{T_-, T_+\}.$$

To test

$$H_o : \text{median of } D = 0 \text{ against } H_a : \text{median of } D < 0$$

reject  $H_o$  if  $T_{obs} = T_+$  with

$$\alpha_{obs} = P(T \leq T_{obs})$$

(otherwise  $\alpha_{obs} \geq 0.5$ ), where  $T$  possesses the Wilcoxon signed rank distribution (Tables A.22-A.25).

To test

$$H_o : \text{median of } D = 0 \text{ against } H_a : \text{median of } D > 0$$

reject  $H_o$  if  $T_{obs} = T_-$  with

$$\alpha_{obs} = P(T \leq T_{obs})$$

(otherwise  $\alpha_{obs} \geq 0.5$ ), where  $T$  possesses the Wilcoxon signed rank distribution (Tables A.22-A.25).

To test

$$H_o : \text{median of } D = 0 \text{ against } H_a : \text{median of } D \neq 0$$

set

$$\alpha_{obs} = 2P(T \leq T_{obs})S$$

where  $T$  possesses the Wilcoxon signed rank distribution (Tables A.22-A.25).

**Example 19.1** In our example, for a two sided test we have

$$T_{obs} = \min\{T_-, T_+\} = \min\{4, 11\} = 4.$$

We have  $n = 5$ , so from the tables we have

$$\alpha_{obs} = 2P(T \leq 4) \approx 0.44,$$

so we do not reject  $H_o$ . ■

For large samples (say,  $n > 12$ ) we can use  $z$ -score

$$Z_{obs} = \frac{T_{obs} - \mu_T}{\sigma_T}$$

where

$$\mu_T = n(n+1)/4, \quad \text{and } \sigma_T = \sqrt{n(n+1)(2n+1)/24},$$

and use approximation  $Z_{obs} \sim N(0, 1)$ .

**Example 19.2** To continue the previous example,

$$\begin{aligned}\mu_T &= n(n+1)/4 = 5 \times 6/4 = 7.5, \\ \sigma_T &= \sqrt{n(n+1)(2n+1)/24} = \sqrt{5 \times 6 \times 11/24} = 3.71\end{aligned}$$

so that

$$Z_{obs} = (4 - 7.5)/3.71 = -0.94$$

giving

$$\alpha_{obs} = 2P(Z \leq Z_{obs}) = 2P(Z \leq -0.94) \approx 2(0.174) = 0.35,$$

roughly the same answer as for the exact procedure. ■

### 19.3 Dealing with Ties in Rank-Based Procedures

Nonparametric methods often require data to be replaced by ranks. For example, the data

$$12, 13, 14, 17, 18, 19, 20 \tag{19.1}$$

would be replaced by ranks

$$1, 2, 3, 4, 5, 6, 7.$$

However, even when the data represents a continuous measurement, such as temperature, ties often occur, as often as not following a rounding off

protocol. While this poses no special problem for methods based on sample means and variances, it does raise a technical problem for rank based procedures.

Suppose the data of Equation (19.1) is replaced by the values:

$$12, 13.5, 13.5, 17, 18.9, 18.9, 18.9. \quad (19.2)$$

It might seem reasonable to assign the ranks 1, 2, 3, 4, 5, 6, 7 as before, so that ranks 5, 6, 7 are used in place of the three common values of 18.9. However, for the signed rank test, and other rank based procedures, the inference depends on various forms of group assignments of the ranks. If in a signed rank test the three values 18.9 occurred at least once in both the '+' and '-' groups, the procedure would depend on how the ranks 5, 6, 7 were allocated.

Clearly, tied values should be treated as ties throughout the procedure, and replaced by the same value. It is also important that an equation such as

$$T_- + T_+ = 1 + 2 + \dots + n = n(n + 1)/2$$

used in the signed rank procedure still hold, so that all rankings still sum to  $n(n + 1)/2$ .

This can be accomplished using the *average rank*. For the data of Equation (19.2) we would use ranks:

$$1, 2.5, 2.5, 4, 6.0, 6.06.0.$$

The largest 3 values from 7 would normally be ranked 5, 6, 7, but if they are tied, they are replaced by average rank  $(5 + 6 + 7)/3 = 6.00$ . This way, the total rank remains  $n(n + 1)/2$ .

**Example 19.3** We may perform a signed rank test on the data of Section 19.1, which contains ties.

Subject	Time after treatment		Change in heart rate	Rank	+/-
	0 mins	30 mins			
1	96	92	-4	5.0	-
2	110	106	-4	5.0	-
3	89	86	-3	2	-
4	95	78	-17	9	-
5	128	124	-4	5.0	-
6	100	98	-2	1	-
7	72	68	-4	5.0	-
8	79	75	-4	5.0	-
9	100	106	+6	8	+

This gives

$$T_{obs} = \min\{T_+, T_-\} = 8,$$

so, for  $n = 9$  we have

$$\alpha_{obs} = 2P(T \leq 8) = 2(0.0489) = 0.0978. \blacksquare$$

## 19.4 Wilcoxon Rank Sum Test

The *Wilcoxon rank sum test* is a commonly used nonparametric alternative to the various two-sample  $t$ -tests, which assume normality of the data. The essential feature of this test is that given two independent samples, these samples are pooled and ranked (ties are assigned average ranks as described above). The values within each sample are replaced by the pooled ranks. The sample with the highest mean or median should also have the highest average rank. The Wilcoxon rank sum test is a method of formally implementing this idea as a formal hypothesis test.

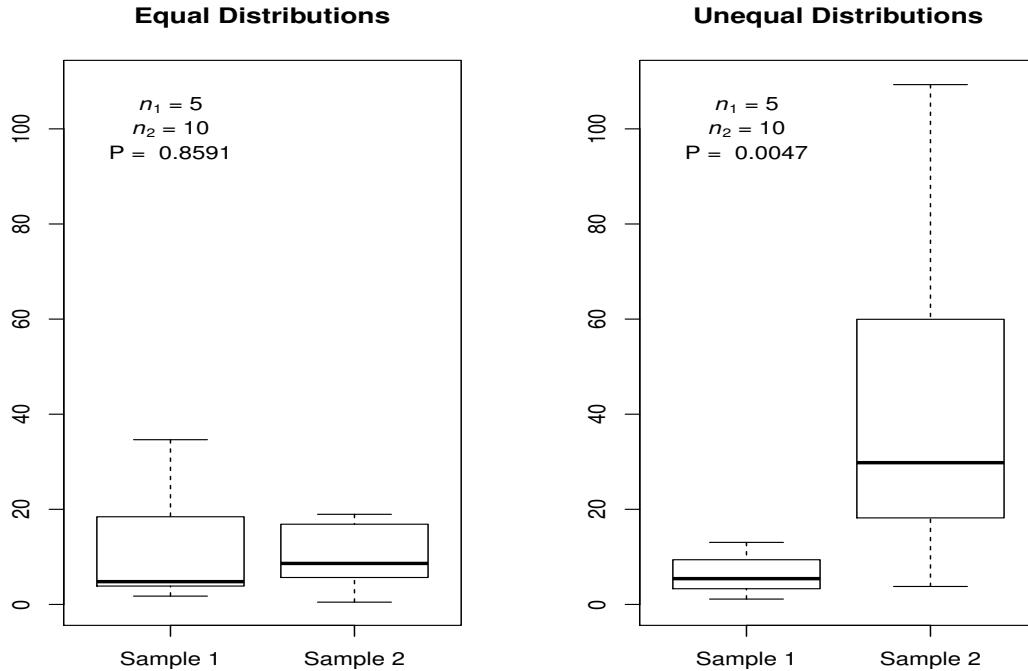


Figure 19.1: Sample data for Wilcoxon rank sum tests.

**Example 19.4** Table 19.1 summarizes two independent samples, with sample sizes  $n_1 = 5, n_2 = 10$ . A two-sample boxplot of the data is given in Figure 19.1 (left plot, titled ‘Equal Distributions’). Despite the relatively small sample size, right-skewness of the data is evident.

Table 19.1: Sample data, two independent samples with pooled ranks. Sample sizes are  $n_1 = 1$  and  $n_2 = 5$ .

Original Values		Ranks	
Sample 1	Sample 2	Sample 1	Sample 2
18.44	9.07	12	9
34.64	0.48	15	1
4.79	5.67	5	6
1.75	7.97	2	7
3.87	8.17	4	8
-	3.15	-	3
-	18.95	-	14
-	16.87	-	11
-	12.08	-	10
-	18.91	-	13

Table 19.2: Pooled ranking procedure for two-sample data in Table 19.1.

Values	Pooled Rank	Sample Group
0.48	1	2
1.75	2	1
3.15	3	2
3.87	4	1
4.79	5	1
5.67	6	2
7.97	7	2
8.17	8	2
9.07	9	2
12.08	10	2
16.87	11	2
18.44	12	1
18.91	13	2
18.95	14	2
34.64	15	1

The pooled ranking procedure assigning the ranks in Table 19.1 is sum-

marized in Table 19.2. Note that the group labels appear to be uniformly distributed among the ranks, as sample group 1 possesses ranks 2, 4, 5, 12, 15. ■

Conventions differ regarding implementation of the rank sum test, however, most approaches can be based on the following. Define

$$\begin{aligned} T_1 &= \text{sum of ranks from sample 1,} \\ T_2 &= \text{sum of ranks from sample 2.} \end{aligned}$$

For the data in Example 19.4 this becomes

$$\begin{aligned} T_1 &= 2 + 4 + 5 + 12 + 15 = 38, \\ T_2 &= 1 + 3 + 6 + 7 + 8 + 9 + 10 + 11 + 13 + 14 = 82. \end{aligned}$$

As in the signed rank test we must have

$$T_1 + T_2 = (n_1 + n_2)(n_1 + n_2 + 1)/2, \quad (19.3)$$

So the test may be carried out with  $T_1$  or  $T_2$  by itself, since

$$T_1 = (n_1 + n_2)(n_1 + n_2 + 1)/2 - T_2.$$

For this reason, it is important that equation (19.3) hold in the presence of ties, which is why tied values should be represented by the appropriate average ranks. For large enough sample sizes (some textbooks recommend  $\min(n_1, n_2) \geq 9$ ) we have approximate normal distributions

$$T_1 \sim N(\mu_1, \sigma_W^2) \text{ and } T_2 \sim N(\mu_2, \sigma_W^2)$$

where

$$\mu_1 = n_1(n_1+n_2+1)/2, \quad \mu_2 = n_2(n_1+n_2+1)/2, \quad \text{and} \quad \sigma_W^2 = n_1n_2(n_1+n_2+1)/12. \quad (19.4)$$

A few remarks should be made. We can compare (19.3) to (19.4) by noting

$$\mu_1 + \mu_2 = (n_1 + n_2)(n_1 + n_2 + 1)/2,$$

that is,  $\mu_1 + \mu_2$  is equal to the total sum of the pooled ranks. Also note that for the two normal distributions in (19.4) the means are not equal but the variances are.

Formally, rank sum tests against the null hypothesis that the medians of each group are equal. Intuitively, if  $T_1$  is larger than expected (equivalently,

$T_2$  is smaller than expected) under the null hypothesis, we have evidence that  $\tilde{\mu}_1 > \tilde{\mu}_2$ , and if  $T_2$  is larger than expected (equivalently,  $T_1$  is smaller than expected) under the null hypothesis, we have evidence that  $\tilde{\mu}_1 < \tilde{\mu}_2$ . We also have one sided and two sided alternatives.

A common convention is to designate the sample group with the smaller sample size as group 1 (tables of critical values often make this assumption). This convention has been followed in Example 19.4.

**One-sided test (lower tail).** To test

$$H_o : \tilde{\mu}_1 \geq \tilde{\mu}_2 \text{ against } H_a : \tilde{\mu}_1 < \tilde{\mu}_2$$

using Tables A.26-A.32 set

$$T_{obs} = T_1$$

and use observed significance level

$$\alpha_{obs} = P(T \leq T_{obs})$$

where  $T$  possesses the Wilcoxon rank sum distribution (Tables A.26-A.32).

To use the normal approximation set, using (19.4),

$$Z_{obs} = \frac{T_1 - \mu_1}{\sigma_W}$$

and use observed significance level

$$\alpha_{obs} = P(Z \leq Z_{obs})$$

where  $Z \sim N(0, 1)$ .

**One-sided test (upper tail).** To test

$$H_o : \tilde{\mu}_1 \leq \tilde{\mu}_2 \text{ against } H_a : \tilde{\mu}_1 > \tilde{\mu}_2$$

using Table A.7 from Pagano and Gavreau set

$$T_{obs} = n_1(n_1 + n_2 + 1) - T_1$$

and use observed significance level

$$\alpha_{obs} = P(T \leq T_{obs})$$

where  $T$  possesses the Wilcoxon rank sum distribution (Tables A.26-A.32).

To use the normal approximation set, using (19.4),

$$Z_{obs} = \frac{T_1 - \mu_1}{\sigma_W}$$

and use observed significance level

$$\alpha_{obs} = P(Z \geq Z_{obs})$$

where  $Z \sim N(0, 1)$ .

**Two-sided test.** To test

$$H_o : \tilde{\mu}_1 = \tilde{\mu}_2 \text{ against } H_a : \tilde{\mu}_1 \neq \tilde{\mu}_2$$

using Table A.7 from Pagano and Gavreau set

$$T_{obs} = \min\{T_1, n_1(n_1 + n_2 + 1) - T_1\}$$

and use observed significance level

$$\alpha_{obs} = 2P(T \leq T_{obs})$$

where  $T$  possesses the Wilcoxon rank sum distribution (Tables A.26-A.32).

To use the normal approximation set, using (19.4),

$$Z_{obs} = \frac{T_1 - \mu_1}{\sigma_W}$$

and use observed significance level

$$\alpha_{obs} = 2P(Z \leq -|Z_{obs}|)$$

where  $Z \sim N(0, 1)$ .

**Example 19.5** To continue with Example (19.4), we had

$$\begin{aligned} T_1 &= 38, \\ T_2 &= 82 \end{aligned}$$

with  $n_1 = 5$ ,  $n_2 = 10$ . If we employ the normal approximation we have

$$\begin{aligned} \mu_1 &= \frac{n_1(n_1 + n_2 + 1)}{2} = \frac{5 \times 16}{2} = 40 \\ \sigma_W^2 &= \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} = \frac{5 \times 10 \times 16}{12} \approx 66.7. \end{aligned}$$

To do a **lower-tailed test**, set

$$T_{obs} = T_1 = 38.$$

From Tables A.26-A.32 we get

$$\alpha_{obs} = P(T \leq 38) \approx 0.4296.$$

Using the normal approximation we have

$$Z_{obs} = \frac{T_1 - \mu_1}{\sigma_W} = \frac{38 - 40}{\sqrt{66.7}} \approx -0.245,$$

giving

$$\alpha_{obs} = P(Z \leq -0.245) \approx 0.40,$$

which is very close to the exact value.

To do an **upper-tailed test**, set

$$T_{obs} = n_1(n_1 + n_2 + 1) - T_1 = 5 \times 80 - 38 = 42$$

The highest available lower tail probability in Tables A.26-A.32 for  $n_1 = 5, n_2 = 10$  is

$$P(T \leq 40) \approx 0.5235,$$

so we conclude  $\alpha_{obs} > 0.5235$ . Using the normal approximation we have

$$\alpha_{obs} = P(Z \geq -0.245) \approx 0.60,$$

which conforms to the exact value.

To do an **two sided test**, set

$$T_{obs} = \min\{n_1(n_1 + n_2 + 1) - T_1, T_1\} = \min\{42, 38\} = 38.$$

From Tables A.26-A.32 we get

$$\alpha_{obs} = 2P(T \leq 38) \approx 0.8592.$$

Using the normal approximation we have

$$\alpha_{obs} = 2P(Z \leq -0.245) \approx 0.80,$$

which is close to the exact value. ■

**Example 19.6** The data in Figure 19.2 are shown as boxplots in Figure 19.1. To do a two sided hypothesis test for equality of medians, first note

$$T_1 = 1 + 2 + 4 + 5 + 6 = 18.$$

The mean and variance of the normal approximation is the same as for Example 19.4

$$\begin{aligned}\mu_1 &= 40 \\ \sigma_W^2 &\approx 66.7.\end{aligned}$$

Table 19.3: Sample data, two independent samples with pooled ranks. Sample sizes are  $n_1 = 1$  and  $n_2 = 5$ .

Original Values		Ranks	
Sample 1	Sample 2	Sample 1	Sample 2
3.3	22.56	2	10
5.43	19.23	4	9
9.4	18.2	5	8
1.13	109.3	1	15
13.05	39.8	6	12
-	75.94	-	14
-	3.78	-	3
-	59.96	-	13
-	16.46	-	7
-	37.06	-	11

We note right away that  $T_1$  is more than 2.5 standard deviations from the mean (that is,  $T_1 - \mu_1 = 22$  and  $\sigma_W \approx 8.2$ ), so we expect that  $H_o$  will be rejected. Formally, we have

$$T_{obs} = \min\{n_1(n_1+n_2+1) - T_1, T_1\} = \min\{80 - 18, 18\} = \min\{62, 18\} = 18,$$

giving  $P$ -value

$$\alpha_{obs} = 2P(T \leq 18) \approx 2 \times 0.0023 = 0.0046.$$

Using the normal approximation we have Using the normal approximation we have

$$Z_{obs} = \frac{T_1 - \mu_1}{\sigma_W} = \frac{18 - 40}{\sqrt{66.7}} \approx -2.69,$$

giving

$$\alpha_{obs} = 2P(Z \leq -2.69) \approx 2 \times 0.004 = 0.008,$$

which is reasonably close to the exact value. ■

## 19.5 Assumptions

For the sign test and signed rank test, the sample consists of independently selected pairs. For the signed rank test, the distribution of the differences under  $H_o$  is assumed to be symmetric about a median of 0.

For the rank sum test each sample is a random sample, and the samples are independent. Under  $H_o$ , the distributions are equal. Note that the

*Mann-Whitney* test is equivalent to the Wilcoxon rank sum test, and so the test is sometimes referred to as the *Mann-Whitney-Wilcoxon* (MWU) test.

The sign test is based on the binomial distribution. To use a normal approximation, minimum sample size guidelines are often suggested (for example  $n > 12$  for the signed rank test,  $\min(n_1, n_2) \geq 9$  for the rank sum test).

# Chapter 20

## Nonparametric Inference in R

R has a number of functions available for nonparametric procedures.

### 20.1 Sign Test

The sign test can be implemented directly from the `pbinom()` function. Consider the following example:

**Example 20.1** This example is from Maskin *et al.*, 1985 in the journal *Circulation*. The following table shows the measurement of heart rate (in beats per minute) before and 30 minutes after administering enalprilat to 9 subjects.

Subject	Time after treatment 0 mins	30 mins	Change in heart rate
1	96	92	-4
2	110	106	-4
3	89	86	-3
4	95	78	-17
5	128	124	-4
6	100	98	-2
7	72	68	-4
8	79	75	-4
9	100	106	+6

We first need to *binarize* the data:

```
> x1 = c(-4, -4, -3, -17, -4, -2, -4, -4, 6)
> x2 = c(x1, 0)
```

```
> x2
[1] -4 -4 -3 -17 -4 -2 -4 -4 6 0
> x3 = x2[x2!=0]
> x3
[1] -4 -4 -3 -17 -4 -2 -4 -4 6
```

The differences were first stored in array `x1`. We need a way of removing ties, so let's add a 0 to `x1`, and store the new data in `x2`. To remove the zero, we can use *array subsetting*. All we need to do is place a conditional statement in the index reference, and the resulting vector will contain only array elements satisfying the condition. The 0 is removed in this way, then the resulting array is stored in `x3`, which is now the same as `x1`.

The following example might clarify this, which shows how we can retain only even numbers within an array of integers:

```
> z1 = c(2,5,4,4,667,775,3330)
> z2 = z1[z1 %% 2 ==0]
> z2
[1] 2 4 4 3330
>
```

To return to the original example, we now want to *binarize* the array `x3`. An array can consist of logical values, for example:

```
> xbin = (x3 > 0)
> xbin
[1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE
```

A condition applied to an array results in a new array of logical outcomes. A logical array can be forced into a binary array simply by multiplying by 1.

```
> xbin = 1*(x3 > 0)
> xbin
[1] 0 0 0 0 0 0 0 0 1
> table(xbin)
xbin
0 1
8 1
> y = sum(xbin)
> y
[1] 1
> n = length(xbin)
> n
[1] 9
>
```

We now have what we need to perform the sign test. The number of positive differences is stored in `y` and the total sample size is stored in `n`. The actual *p*-value is now obtained from the binomial CDF function `pbinom()`.

```
> pval = 2*pbinom(min(y, n-y), prob = 0.5, size = n)
> pval
[1] 0.0390625
>
```

We therefore have a *p*-value  $\leq 0.04$  for rejecting the null hypothesis  $H_0 : p = 1/2$  against  $H_a p \neq 1/2$ , where  $p$  is the probability that the change in heart rate is positive for a given subject. ■

Note that an actual sign test function might be available from a repository. There exists, for example a package called `BDSA` which includes a sign test function `sign.test()`, which can, in principle, be installed using the `install.packages()` function (see Lecture 3). Unfortunately, I got the following result:

```
> install.packages("BDSA")
Warning message:
package ‘BDSA’ is not available (for \R version 3.0.1)
>
```

A good practice is to keep an older version of `R` on your computer, for cases like this.

## 20.2 Signed Rank and Rank Sum Procedures

Recall that the `t.test()` function can be used for both paired and independent sample t-tests. Similarly, the function `wilcox.test()` is used for both the signed rank and rank sum test by using the `paired` option in the same way. In addition, the distributions for the respective test statistics are also available using the `psignrank()` and `pwilcox()` functions. For example, from Example 18.1 we had a signed rank observed statistic of value  $T_{obs} = 4$ . For a sample of size  $n = 11$  we can easily get the *p*-value

```
> 2*psignrank(4,n=5)
[1] 0.4375
```

We can also use the `wilcox.test()` with option `paired = T`. One slight technical issue arises. With the paired option, `wilcox.test()` expects two samples. If we start with the differences, we can ‘fool’ the function by creating a second sample of zeros, which would be mathematically equivalent:

```

> x = c(-3.4, -1.4, 3.2, 4.3, 5.9)
> y = rep(0, length(x))
> y
[1] 0 0 0 0 0
> wilcox.test(x, y, paired=T)

Wilcoxon signed rank test

data: x and y
V = 11, p-value = 0.4375
alternative hypothesis: true location shift is not equal to 0

```

>

which gives the same result.

There is an important issue relating to conventions when using the rank sum procedure. The smallest value possible for  $T_1$  is

$$\min T_1 = 1 + \dots + n_1 = n_1(n_1 + 1)/2.$$

It is often the practice to subtract this minimum from the rank sum, in which case the values range from 0 to  $n_1 n_2$ , so the test statistic becomes

$$T_{obs} = T_1 - n_1(n_1 + 1)/2.$$

In fact, the functions `pwilcox()` and `wilcox.test()` both adopt this convention. For Example 18.4 this would yield test statistic

$$T_{obs} = 38 - 5 \times 6/2 = 23,$$

so we would use this value. If we have  $T_{obs}$  we can get the  $p$ -value directly from the rank sum distribution function `p(wilcox)`:

```

> 2*pwilcox(23, 5, 10)
[1] 0.8591409
>

```

which gives (almost) the same value as that obtained from the tables in Example 18.5. Note that we need to specify both sample sizes  $n_1, n_2$ .

We can also input the data directly, either entering the two samples separately, or using the model notation used earlier for the `aov()` command in Lecture 17.

```

> y1 = c(18.44, 34.64, 4.79, 1.75, 3.87)
> y2 = c(9.07, 0.48, 5.67, 7.97, 8.17,

```

```
3.15, 18.95, 16.87, 12.08, 18.91)
> y = c(y1, y2)
> x = c(rep(0,5), rep(1,10))
> cbind(y,x)
   y x
[1,] 18.44 0
[2,] 34.64 0
[3,]  4.79 0
[4,]  1.75 0
[5,]  3.87 0
[6,]  9.07 1
[7,]  0.48 1
[8,]  5.67 1
[9,]  7.97 1
[10,]  8.17 1
[11,]  3.15 1
[12,] 18.95 1
[13,] 16.87 1
[14,] 12.08 1
[15,] 18.91 1
> wilcox.test(y1,y2)

Wilcoxon rank sum test

data: y1 and y2
W = 23, p-value = 0.8591
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(y ~ x)

Wilcoxon rank sum test

data: y by x
W = 23, p-value = 0.8591
alternative hypothesis: true location shift is not equal to 0

>
```

# Chapter 21

## Inference for Correlation

We have seen in Section 11.3.2 the sample correlation coefficient as a means of assigning a measure to the degree of association between data consisting of paired observations. Suppose measurements of two variables are given by

$$X_1, X_2, \dots, X_n$$

and

$$Y_1, Y_2, \dots, Y_n.$$

The sample correlation coefficient was defined by

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}.$$

In fact, this particular definition of  $r$  is referred to more specifically as the *Pearson product-moment correlation coefficient* and we will see other types of correlation coefficients below. However, the usual convention is that when reference is made to the *correlation coefficient*, the Pearson correlation is intended. Just as the various statistics  $\bar{X}$  and  $S^2$  estimate population parameters  $\mu$  and  $\sigma^2$ , the sample correlation  $r$  estimates a population correlation

$$\rho = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y},$$

where  $\mu_X, \mu_Y$  and  $\sigma_X^2, \sigma_Y^2$  the the respective means and variances of random variables  $X$  and  $Y$ .

### 21.1 Inference for Correlations

Consider the test statistic

$$T = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}, \quad (21.1)$$

where  $n$  is the number of paired observations. If  $X$  and  $Y$  together possess a bivariate normal distribution (we will define this precisely in Section 25.3) then  $T$  possesses a  $t$ -distribution with  $n - 2$  degrees of freedom under the null hypothesis

$$H_0 : \rho = 0,$$

so a lower tailed, upper tailed or two-sided  $t$ -test may be used for testing against this hypothesis. A critical value for  $r$  may be obtained from the  $t$ -distribution usign the inverse transformation of (21.1):

$$r = \frac{T}{\sqrt{n - 2 + T^2}}, \quad (21.2)$$

If the bivariate normal assumption does not hold, the  $t$ -distribution holds approximately as long as  $n$  is large, or if  $n$  is small and neither  $X$  or  $Y$  deviate greatly from normality, which can be checked with, for example, a quantile plot.

**Example 21.1** The following example is reported in Devore (Example 12.10, *Probability and Statistics for Engineering and the Sciences, Fourth Edition*). It represent  $n = 11$  paired measurements of rubber tensile modulus ( $y$ ) and percent bound-rubber content ( $x$ ).

X = 16.1 31.5 21.5 22.4 20.5 28.4 30.3 25.6 32.7 29.2 34.7  
Y = 4.41 6.81 5.26 5.99 5.92 6.14 6.84 5.87 7.03 6.89 7.87

Normal quantile plots, and a scatter plot are shown in Figure 21.1, and the assumption of bivariate normality appears to hold. The sample correlation coefficient is

$$r_{XY} = 0.939,$$

which is a very high value. The t-test against  $H_0 : \rho_{XY} = 0$  is carried out using statistic

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.939\sqrt{11-2}}{\sqrt{1-0.939^2}} \approx 8.19.$$

This is clearly a very large value,  $T_{obs} > t_{9,0.0005} = 4.781$ , so the null hypothesis is easily rejected.

If we wish to express the rejection region directly in terms of the correlation, we can apply the inverse transformation (21.2) directly to any critical value from the  $t$ -distribution. For example, if we wish to construct a level  $\alpha = 0.01$  critical region applicable directly to  $r$ , we use, given  $t_{9,0.005} = 3.2498$ ,

$$r_{\alpha/2} = \frac{t_{\alpha/2}}{\sqrt{n-2+t_{\alpha/2}^2}} = \frac{3.2498}{\sqrt{9+3.2498^2}} = 0.735,$$

that is, we reject  $H_o : \rho_{XY} = 0$  in favor of the two-sided alternative hypothesis at a  $\sigma = 0.01$  significance level if  $|r| \geq 0.735$  (which we do with our given data). ■

## 21.2 Sample Size Analysis for Hypothesis Tests

We may also use Equation (21.2) to answer a quite simple, but important, question: given a sample size  $n$ , how large does a sample correlation coefficient have to be in order to be reasonably confident that the true correlation  $\rho$  is not 0? We may express this question as a two sided hypothesis test against  $H_o : \rho = 0$ , which we reject with significance level  $\alpha$  if  $|T| \geq t_{n-2,\alpha/2}$ , for  $T$  defined by (21.1), or if  $|r| \geq r_{\alpha/2}$ , where

$$r_{\alpha/2} = \frac{t_{\alpha/2}}{\sqrt{n-2+t_{\alpha/2}^2}},$$

following inverse transformation (21.2). Values of  $r_{\alpha/2}$  for varying sample sizes  $n$  are shown in Figure 21.2.

## 21.3 Inference for the Pearson Correlation Coefficient When $\rho \neq 0$

The distribution of  $r$  can be given exactly under the assumption of bivariate normality, when  $\rho = 0$ . When  $\rho \neq 0$  a normal approximation can be induced by the *Fisher transformation*:

$$V = \frac{1}{2} \ln \left( \frac{1+r}{1-r} \right). \quad (21.3)$$

It can be shown that the following normal approximation holds:

$$V \sim N(\mu_{V,\rho}, \sigma_{\rho}^2), \quad \text{where } \mu_{V,\rho} = \frac{1}{2} \ln \left( \frac{1+\rho}{1-\rho} \right) \text{ and } \sigma_{\rho}^2 = \frac{1}{n-3}. \quad (21.4)$$

Note that the transformations defining  $V$  and  $\mu_{V,\rho}$  are increasing functions of  $r$  and  $\rho$  respectively, and so the Fisher transformation may be used in much the same way the transformation (21.1) was used. For example, a test against null hypothesis

$$H_o : \rho = \rho_0$$

for some hypothetical value  $\rho_0$  can be based on test statistic

$$Z = \frac{V - \mu_{V,\rho_0}}{\sqrt{\frac{1}{n-3}}},$$

which would have approximately a standard normal distribution  $N(0, 1)$  under  $H_0$ . Because the Fisher transformation holds for any  $\rho$ , a confidence interval can be obtained for  $\rho$ . The first step is to obtain a level  $1 - \alpha$  confidence interval for  $\mu_{V,\rho}$ , which by (21.4) is easily seen to be, for sample size  $n$  and observed correlation coefficient  $r_{obs}$ :

$$\left( V_{obs} - \frac{z_{\alpha/2}}{\sqrt{n-3}}, V_{obs} + \frac{z_{\alpha/2}}{\sqrt{n-3}} \right) = (L_V, U_V), \quad (21.5)$$

where

$$V_{obs} = \frac{1}{2} \ln \left( \frac{1+r_{obs}}{1-r_{obs}} \right),$$

and  $z_{\alpha/2}$  is a critical value from  $N(0, 1)$ . We then invert the expression for  $\mu_{V,\rho}$ , giving a direct level  $1 - \alpha$  confidence interval for  $\rho$  of the form

$$\left( \frac{e^{2L_V} - 1}{e^{2L_V} + 1}, \frac{e^{2U_V} - 1}{e^{2U_V} + 1} \right). \quad (21.6)$$

**Example 21.2** In Example 23.1 we had  $r_{XY} = 0.939$  with sample size  $n = 11$ . Suppose we want a 95% confidence interval for  $\rho$ . We have transformation

$$\begin{aligned} V_{obs} &= \frac{1}{2} \ln \left( \frac{1+r_{obs}}{1-r_{obs}} \right) \\ &= \frac{1}{2} \ln \left( \frac{1+0.939}{1-0.939} \right) = 1.7295, \end{aligned}$$

so we first use the confidence interval for  $\mu_{V,\rho}$  given in (21.5):

$$\begin{aligned} \left( V_{obs} - \frac{z_{\alpha/2}}{\sqrt{n-3}}, V_{obs} + \frac{z_{\alpha/2}}{\sqrt{n-3}} \right) &= \left( 1.7295 - \frac{1.96}{\sqrt{8}}, 1.7295 + \frac{1.96}{\sqrt{8}} \right) \\ &= (1.0365, 2.4225), \end{aligned}$$

giving  $L_V = 1.0365$ ,  $U_V = 2.4225$ . The confidence interval for  $\rho$  is then directly given by (21.6), after direct substitution of  $L_V$  and  $U_V$ :

$$\left( \frac{e^{2L_V} - 1}{e^{2L_V} + 1}, \frac{e^{2U_V} - 1}{e^{2U_V} + 1} \right) = (0.7765, 0.9844). \quad \blacksquare$$

## 21.4 Nonparametric Correlation Coefficients

The Pearson correlation coefficient is sensitive to outliers, and spuriously large values can occur when the underlying assumptions do not hold.

The *Spearman rank correlation*  $r_S$  is a rank-based nonparametric alternative. The procedure is straightforward. The values of  $X$  and  $Y$  are

replaced by their ranks (each is ranked separately). Then  $r_S$  is simply the Pearson correlation coefficient of the ranks. For large enough  $n$ , the test procedure used for the Pearson correlation coefficient is a good approximation. Note that  $r_S$  has the same quantitative interpretation as  $r$ . Ties may be handled using the average rank method of Section 19.3.

Another commonly used alternative is the *Kendall rank correlation coefficient*, or *Kendall's  $\tau$* . If bivariate data  $(X_i, Y_i)$  is in *positive concordance*, then for any two indicies  $i, j$  there will a tendency to observe  $X_i > X_j$  when  $Y_i > Y_j$ , and  $X_i < X_j$  when  $Y_i < Y_j$ . In either case, we would have  $(X_i - X_j)(Y_i - Y_j) > 0$ . We would expect to see this, for example, with positively correlated data. Conversely, with *negative concordance*, for any two indicies  $i, j$  there will a tendency to observe  $X_i < X_j$  when  $Y_i > Y_j$ , and  $X_i > X_j$  when  $Y_i < Y_j$ . In either case, we would have  $(X_i - X_j)(Y_i - Y_j) < 0$ .

To calculate the Kendall rank correlation coefficient, we enumerate every unordered pair of indicies  $i, j$ , of which there are  $n(n - 1)/2$ . Then define

$$\kappa = \sum_{i,j} \kappa_{i,j}$$

where  $\kappa_{i,j} = 1$  if the pairs  $(X_i, Y_i)$  and  $(X_j, Y_j)$  are in positive concordance  $((X_i - X_j)(Y_i - Y_j) > 0)$ ;  $\kappa_{i,j} = -1$  if the pairs  $(X_i, Y_i)$  and  $(X_j, Y_j)$  are in negative concordance  $((X_i - X_j)(Y_i - Y_j) < 0)$ ; and  $\kappa_{i,j} = 0$  if  $(X_i - X_j)(Y_i - Y_j) = 0$ . The coefficient (Kendall's  $\tau$ ) is then:

$$\tau = \frac{\kappa}{n(n - 1)/2}.$$

There are various approximation methods for the distributions of the Spearman and Kendall coefficients, which are generally combinatoric in nature, and require computer algorithms to evaluate.

## 21.5 Inference for Correlations in R

Inference for correlation in R may be done using function `cor.test()`. The function `cor()` calculates correlations, but does not perform inference. The options are given below:

```
cor.test(x, y,
         alternative = c("two.sided", "less", "greater"),
         method = c("pearson", "kendall", "spearman"),
         exact = NULL, conf.level = 0.95,
         continuity = FALSE, ...)
```

The values `x`, `y` represent the paired samples, and must be numerical vectors of equal length. The `method` option specifies the Pearson, Spearman

or Kendall coefficients discussed above. The default is the Pearson coefficient. Confidence intervals are available only for the Pearson coefficients. Otherwise, tests against the hypothesis  $H_0 : \rho = 0$  are available.

To continue from Example 23.2 we apply `cor.test()` to the data, for each of the three coefficients discussed:

```
>
> X = c(16.1, 31.5, 21.5, 22.4, 20.5, 28.4, 30.3, 25.6, 32.7, 29.2, 34.7)
> Y = c(4.41, 6.81, 5.26, 5.99, 5.92, 6.14, 6.84, 5.87, 7.03, 6.89, 7.87)
>
> cor(X,Y,method=c("pearson"))
[1] 0.9388037
> cor(X,Y,method=c("spearman"))
[1] 0.9181818
> cor(X,Y,method=c("kendall"))
[1] 0.7818182
>
>
> cor.test(X,Y,method=c("pearson"))

Pearson's product-moment correlation

data: X and Y
t = 8.1765, df = 9, p-value = 1.859e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.7758731 0.9843355
sample estimates:
cor
0.9388037

> cor.test(X,Y,method=c("spearman"))

Spearman's rank correlation rho

data: X and Y
S = 18, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.9181818

> cor.test(X,Y,method=c("kendall"))
```

```
Kendall's rank correlation tau

data: X and Y
T = 49, p-value = 0.0003334
alternative hypothesis: true tau is not equal to 0
sample estimates:
tau
0.7818182

>
```

We obtain the same confidence interval seen in Example 23.2. We also get coefficients  $r = 0.939$ ,  $r_S = 0.918$ ,  $\tau = 0.782$ . Generally, these coefficients will agree. The use of  $\tau$  is especially appropriate when  $X$  and  $Y$  are positively or negatively associated, but not linearly associated.

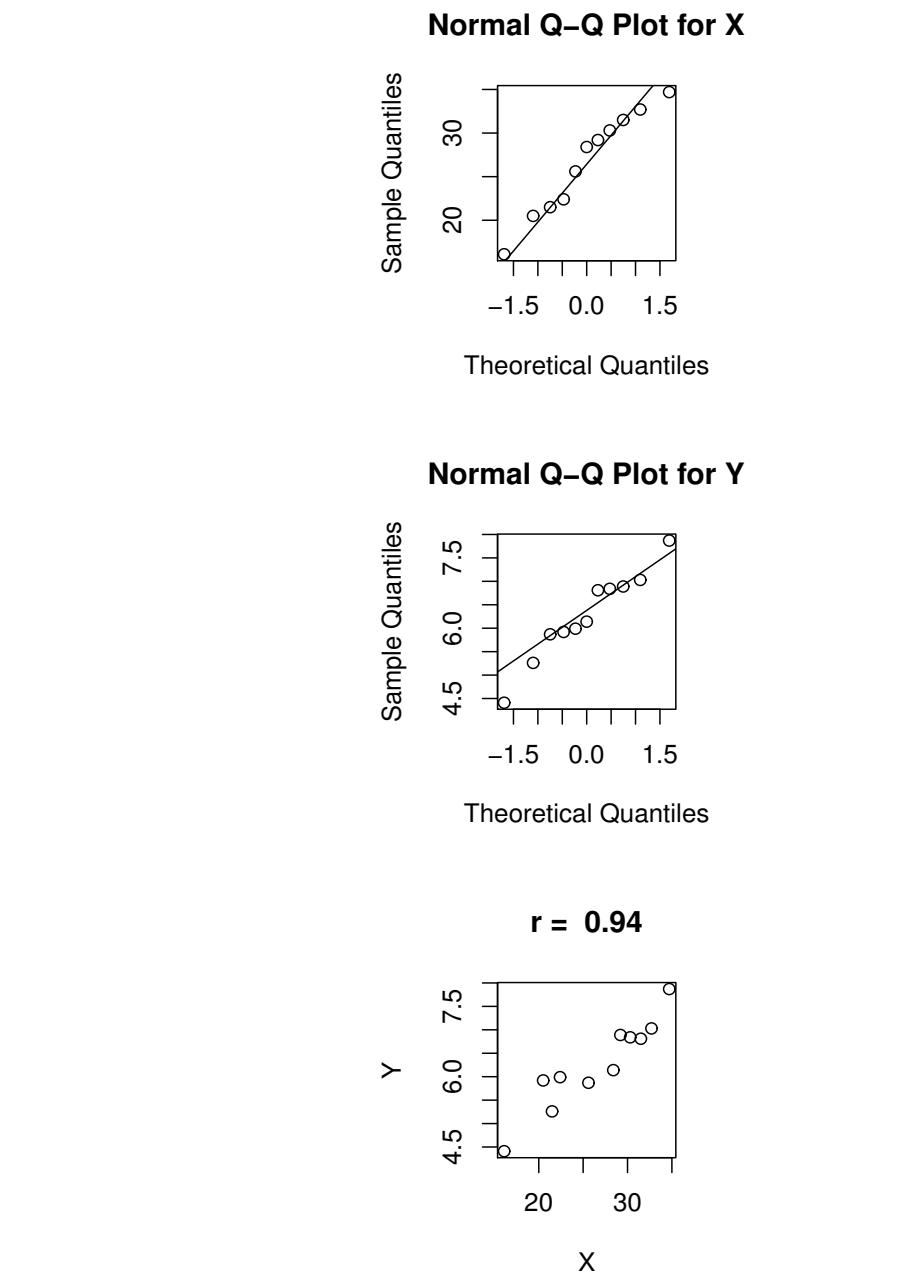


Figure 21.1: Normal quantile plots and scatter plot for Example 23.1

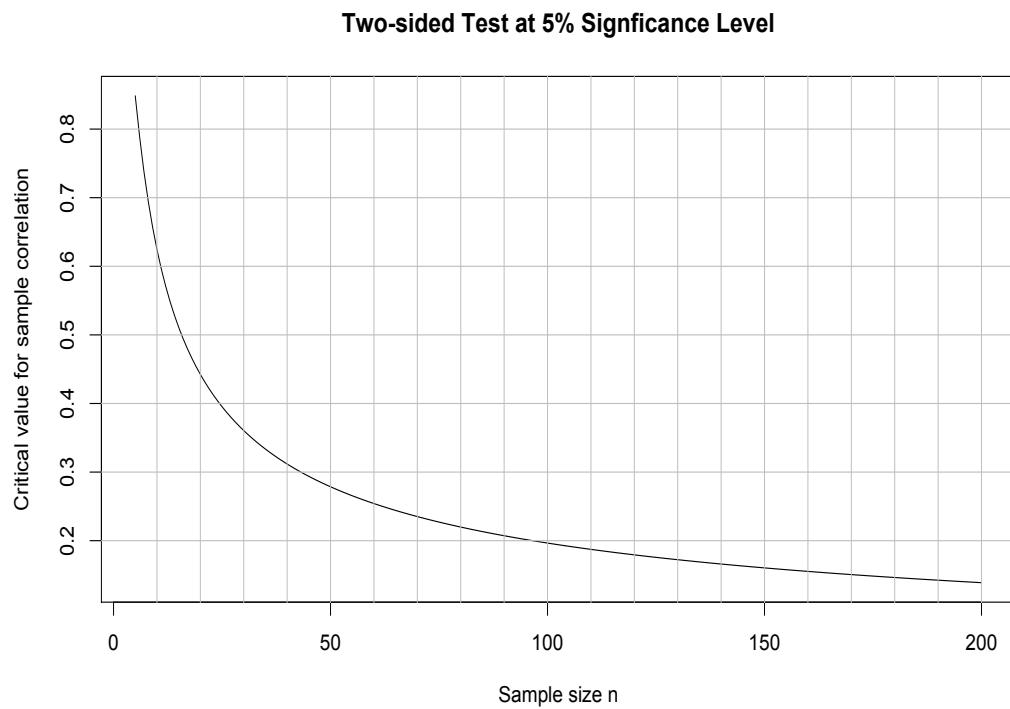


Figure 21.2: Values of  $r_{\alpha/2}$  for varying sample sizes  $n$ . This gives the critical value for a two-sided test against  $H_o : \rho = 0$  for significance level  $\alpha = 0.05$ .

## Chapter 22

# Goodness of Fit Tests and Contingency Tables

Suppose we collect a sample of size  $n$  of categorical data, consisting of  $r$  categories. This will sometimes be numerical data reduced to categories, as is done when constructing a histogram.

We hypothesize that these categories exists in the population according to frequencies  $p_1, \dots, p_r$ , and so we wish to test the hypotheses

$H_o$  :  $p_1, \dots, p_r$  are the population frequencies for categories  $1, \dots, r$

$H_a$  : At least one of the hypothetical frequencies is incorrect

The statistic we use is given by

$$X^2 = \sum_{i=1}^r \frac{(O_i - E_i)^2}{E_i}$$

where

$$\begin{aligned} O_i &= \text{Observed count for category } i \\ &= n_i \\ E_i &= \text{Expected count for category } i \\ &= np_i \end{aligned}$$

It can be shown that if  $H_o$  is true, then  $X^2$  has approximately a  $\chi^2$  distribution with  $r - 1$  degrees of freedom. Critical values for this distribution are given in Table A.4. The form is identical to the  $t$ -distribution tables. This means the observed level of significance is

$$\alpha_{obs} = P(\chi_\nu^2 > X^2)$$

where  $\chi_{\nu}^2$  is a random variable with a  $\chi^2$  distribution with  $\nu = r - 1$  degrees of freedom. A size  $\alpha$  rejection region is given by

$$X^2 \geq \chi_{r-1,\alpha}^2$$

where  $\chi_{r-1,\alpha}^2$  is the  $\alpha$  critical value for a  $\chi^2$  distribution with  $r - 1$  degrees of freedom.

**Example 22.1** In order to test whether or not a dice is loaded it is tossed 600 times, and the frequency of each outcome is recorded. If the probability of tossing outcome  $i$  is  $p_i$  our hypotheses are

$$H_o : p_i = 1/6, \quad i = 1, \dots, 6$$

$$H_a : p_i \neq 1/6 \text{ for at least one } i$$

Suppose the observed frequencies (listed with the expected frequencies) are given below.

Outcome	Observed Frequency	Expected Frequency
1	$n_1 = 79$	$np_1 = 100$
2	$n_2 = 99$	$np_2 = 100$
3	$n_3 = 110$	$np_3 = 100$
4	$n_4 = 91$	$np_4 = 100$
5	$n_5 = 96$	$np_5 = 100$
6	$n_6 = 125$	$np_6 = 100$

The test statistic is then

$$\begin{aligned} X^2 &= \sum_{i=1}^6 \frac{(n_i - np_i)^2}{np_i} \\ &= \frac{(79 - 100)^2}{100} + \frac{(99 - 100)^2}{100} + \frac{(110 - 100)^2}{100} \\ &\quad + \frac{(91 - 100)^2}{100} + \frac{(96 - 100)^2}{100} + \frac{(125 - 100)^2}{100} \\ &= 12.64 \end{aligned}$$

Since  $r = 6$  the degrees of freedom is  $r - 1 = 5$ . If we wish to test at a 0.05% significance level, we obtain from the tables,

$$\chi_{5,0.05}^2 = 11.07$$

where we use  $r - 1 = 5$  degrees of freedom. Since

$$X^2 > \chi_{5,0.05}^2$$

we may reject the null hypothesis of equal frequencies at a 5% significance level. ■

## 22.1 Yates's Correction

The use of the  $\chi^2$  distribution to model count data is comparable to the use of the normal distribution to model a binomial random variable. A correction method for this approximation was discussed in Section 8.4.1. This type of correction can also be introduced into the goodness of fit statistic using *Yates's correction for continuity*:

$$X_{\text{Yates}}^2 = \sum_{i=1}^r \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

Notice that for each term, the effect of introducing the correction is to decrease its' contribution to the statistic (unless  $|O_i - E_i| \leq 0.25$ ). Yates's correction is therefore a conservative procedure, which prevents approximation error from unduly influencing the P-value in a downward direction.

## 22.2 Assumptions

To use the technique introduced here we assume that the sample selected is a true random sample. The  $\chi^2$  distribution for  $X^2$  is an approximation which is useful provided that the count in each cell is large enough. Many statistical computer packages will warn the user if any cell contains a count less than 5, which is a commonly used rule of thumb. Modifications of the technique presented here exist which are designed to correct for small cell counts. Yates' correction, described above is the most common, and many computer packages give this as an option.

## 22.3 Hypothesis Tests for Contingency Tables

Recall the "caffiene consumption" example (Example 11.1):

Marital status	Caffeine Consumption (mg/day)				Total
	0	1-150	151-300	> 300	
Married	652	1537	598	242	3029
Divorced, separated or widowed	36	46	38	21	141
Single	218	327	106	67	718
Total	906	1910	742	330	3888

A central question was whether or not there was any interaction between marital status and the amount of caffeine consumed. It appeared that those

subjects in the divorced category consumed more caffeine. However, to reach such a conclusion we need to be able to assess the amount of statistical evidence available in the contingency table to support it. Otherwise, we would not be able to say definitively that the effect we see is not simply attributable to random variation that occurs with all sampling.

We therefore need a suitable hypothesis test. To construct one we rely on the concept of independence introduced in Section 5.6. Suppose a new subject is randomly selected from the population from which the original sample was drawn. Define the two events

$$\begin{aligned} A &= \{ \text{Subject is divorced, separated or widowed} \} \\ B &= \{ \text{Subject drinks more than 300 mg/day} \} \end{aligned}$$

If there is no interaction between marital status and caffeine consumption then we would expect the events  $A$  and  $B$  to be independent. From the contingency tables we can estimate

$$\begin{aligned} P(A) &\approx 141/3888 \\ P(B) &\approx 330/3888 \end{aligned}$$

from the row and column totals. In addition we can estimate the probability

$$P(A \cap B) \approx 21/3888$$

since there are 21 subjects in the cell corresponding to  $A$  and  $B$ . If  $A$  and  $B$  are independent we expect that

$$P(A \cap B) \approx P(A)P(B).$$

According to the estimates we have

$$\begin{aligned} P(A)P(B) &\approx \frac{141}{3888} \times \frac{330}{3888} \\ &= 0.00308 \\ P(A \cap B) &\approx \frac{21}{3888} \\ &= 0.00540. \end{aligned}$$

The two probabilities are not the same. But clearly, even if  $A$  and  $B$  were independent we would not expect the two probabilities to be exactly the same, since there will always be some random sampling variation.

The answer is to set up a hypothesis  $H_o$  that all row events are independent of all column events. Then construct a test statistic whose distribution is known when it is assumed that  $H_o$  is true. Calculate the statistic, and if it is compatible with  $H_o$ , then we do not reject the hypothesis  $H_o$  of row/column independence.

Let

$$\begin{aligned} r_i &= P(\text{ith row event occurs}) \\ c_j &= P(\text{jth column event occurs}) \\ p_{ij} &= P(\text{ith row event AND jth column event occur}) \end{aligned}$$

If the  $i$ th row event and the  $j$ th column event are independent then

$$p_{ij} = r_i c_j$$

so that the hypothesis of row/column independence can be written

$$\begin{aligned} H_0 : p_{ij} &= r_i c_j \text{ for all row events } i \text{ and column events } j \\ H_a : p_{ij} &\neq r_i c_j \text{ for some row event } i \text{ and column event } j. \end{aligned}$$

The statistic we use is given by

$$X^2 = \sum_i^{n_r} \sum_j^{n_c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where

$$\begin{aligned} O_{ij} &= \text{Observed count in cell } i, j \\ E_{ij} &= \text{Expected count in cell } i, j \text{ under } H_0 \end{aligned}$$

and

$$\begin{aligned} n_r &= \text{The number of rows} \\ n_c &= \text{The number of columns.} \end{aligned}$$

The quantity  $O_{ij}$  is simply the count given in the cell given by row  $i$  and column  $j$ . To obtain  $E_{ij}$  we note that under  $H_0$  we have  $p_{ij} = r_i c_j$ . We may interpret  $p_{ij}$  as the proportion of the total sample in cell  $i, j$ . If we let the total sample size be  $N$ , then if  $H_0$  is true we should have

$$\begin{aligned} E_{ij} &= N p_{ij} \\ &= N r_i c_j. \end{aligned}$$

If we let

$$\begin{aligned} R_i &= \text{Total counts in row } i \\ C_j &= \text{Total counts in column } j \end{aligned}$$

then as an approximation we have

$$\begin{aligned} r_i &= \frac{R_i}{N} \\ c_j &= \frac{C_j}{N} \end{aligned}$$

which gives, approximately

$$\begin{aligned} E_{ij} &= N \frac{R_i}{N} \frac{C_j}{N} \\ &= \frac{R_i C_j}{N}. \end{aligned}$$

If  $H_o$  is not true then we would expect  $O_{ij}$  to be significantly different from  $E_{ij}$  for at least some cells, which in turn would tend to make the test statistic  $X^2$  large.

It can be shown that if  $H_o$  is true, then  $X^2$  has approximately a  $\chi^2$  distribution with  $(n_r - 1)(n_c - 1)$  degrees of freedom. This means the observed level of significance is

$$\alpha_{obs} = P(\chi_\nu^2 > X^2)$$

where  $\chi_\nu^2$  is a random variable with a  $\chi^2$  distribution with  $\nu = (n_r - 1)(n_c - 1)$  degrees of freedom. A size  $\alpha$  rejection region is given by

$$X^2 \geq \chi_{(n_r-1)(n_c-1),\alpha}^2$$

where  $\chi_{(n_r-1)(n_c-1),\alpha}^2$  is the  $\alpha$  critical value for a  $\chi^2$  distribution with  $(n_r - 1)(n_c - 1)$  degrees of freedom.

**Example 22.2** To continue with the caffeine consumption example, we need to calculate for each cell the expected count under the hypothesis

$$H_o : p_{ij} = r_i c_j \text{ for all row events } i \text{ and column events } j.$$

As an example, if we take the cell in row  $i = 1$  and column  $j = 1$  then the required row and column totals are

$$\begin{aligned} R_1 &= 3029 \\ C_1 &= 906 \end{aligned}$$

with total count

$$N = 3888$$

so that the expected count for cell 1,1 is

$$\begin{aligned} E_{11} &= \frac{R_1 C_1}{N} \\ &= \frac{3029 \times 906}{3888} \\ &= 705.8 \end{aligned}$$

and the contribution to the  $X^2$  statistic is

$$\begin{aligned} X_{ij}^2 &= \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \\ &= \frac{(652 - 705.8)^2}{705.8} \\ &= 4.11 \end{aligned}$$

Then  $X^2$  is calculated by summing the contribution  $X_{ij}^2$  for each cell. The following table gives for each cell the observed count  $O_{ij}$ , the expected count  $E_{ij}$  and the contribution  $X_{ij}^2$  to the test statistic  $X^2$ , along with row and column totals for each quantity. (Note that the row and column totals for  $O_{ij}$  and  $E_{ij}$  are the same. Why is this?).

Marital status	Caffeine Consumption (mg/day)				Total
	0	1-150	151-300	> 300	
Married	$O_{11} = 652$	$O_{12} = 1537$	$O_{13} = 598$	$O_{14} = 242$	3029
	$E_{11} = 705.8$	$E_{12} = 1488.0$	$E_{13} = 578.1$	$E_{14} = 257.1$	3029
	$X_{11}^2 = 4.11$	$X_{12}^2 = 1.61$	$X_{13}^2 = 0.69$	$X_{14}^2 = 0.89$	7.30
Divorced, separated or widowed	$O_{21} = 36$	$O_{22} = 46$	$O_{23} = 38$	$O_{24} = 21$	141
	$E_{21} = 32.9$	$E_{22} = 69.3$	$E_{23} = 26.9$	$E_{24} = 12.0$	141
	$X_{21}^2 = 0.30$	$X_{22}^2 = 7.82$	$X_{23}^2 = 4.57$	$X_{24}^2 = 6.82$	19.51
Single	$O_{31} = 218$	$O_{32} = 327$	$O_{33} = 106$	$O_{34} = 67$	718
	$E_{31} = 167.3$	$E_{32} = 352.7$	$E_{33} = 137.0$	$E_{34} = 60.9$	7.18
	$X_{31}^2 = 15.36$	$X_{32}^2 = 1.88$	$X_{33}^2 = 7.02$	$X_{34}^2 = 0.60$	24.86
Total	906	1910	742	330	3888
	906	1910	742	330	3888
	19.77	11.31	12.28	8.31	51.66

If we wish to test at a 0.05% significance level, we obtain from the tables,

$$\chi^2_{6,0.05} = 12.59$$

where we use  $(n_r - 1)(n_c - 1) = 6$  degrees of freedom. If  $X^2$  exceeds 12.59 we may reject  $H_0$  at a 5% significance level. Note that there is one single cell, in row 3, column 1 whose contribution  $X_{31}^2 = 15.36$  by itself exceeds that critical value, so it is actually unnecessary to do the complete sum to know that

$$X^2 > \chi^2_{6,0.05}$$

so that we may reject the hypothesis  $H_0$  of row/column independence at a 5% significance level.

To find the observed significance level, we see that the row corresponding to 6 degrees of freedom has as its' largest value

$$\chi^2_{6,0.005} = 18.55.$$

The total sum of the contributions is, (after some math)

$$\begin{aligned} X^2 &= 51.66 \\ &> \chi^2_{6,0.005} \end{aligned}$$

so that the observed significance level

$$\alpha_{obs} < 0.005$$

indicating that the evidence with which  $H_0$  is rejected is very strong. ■

## 22.4 Yates's Correction

Yates's correction for continuity, discussed in Section 22.1 applies also to the test for independence in a contingency table, in particular,

$$X^2_{Yates} = \sum_i^{n_r} \sum_j^{n_c} \frac{(|O_{ij} - E_{ij}| - 0.5)^2}{E_{ij}}$$

where the observed and expected counts  $O_{ij}$  and  $E_{ij}$  are given above.

## 22.5 $\chi^2$ Tests in R

$\chi^2$  Tests in R are supported by the `chisq.test()` function:

```
chisq.test(x, y = NULL, correct = TRUE,
           p = rep(1/length(x), length(x)), rescale.p = FALSE,
           simulate.p.value = FALSE, B = 2000)
```

For goodness of fit test, the object `x` is a vector of counts, and `p` is a vector of hypothetical probabilities. Note that the default is equal probabilities, appropriate for testing the fairness of dice. There are two methods for testing independence in a contingency table. The first is to set `x` to be the matrix of counts. In this case `y` is not needed. Otherwise `x` and `y` are factors representing row and column categories, respectively. Each case is represented once in each factor. This would be appropriate when constructing a contingency table directly from a data frame. Yate's correction is implemented by default for  $2 \times 2$ , but is otherwise not used.

The following example illustrates the goodness of fit test used to compare a set of 100 integer measurements ranging from 0 to 5 to a  $bin(100, 0.5)$  distribution, making use of R's simulation utilities.

```

> pr = dbinom(0:5,size=5,prob=0.5)
> x = rbinom(n=100,size=5,prob=0.5)
> table(x)
x
0 1 2 3 4 5
2 23 29 31 10 5
> chisq.test(table(x),p=pr)

Chi-squared test for given probabilities

data: table(x)
X-squared = 7.2, df = 5, p-value = 0.2062

Warning message:
In chisq.test(table(x), p = pr) :
  Chi-squared approximation may be incorrect
>

```

The  $P$ -value is 0.2062, so that the data is compatible with the hypothesis, as we expect. Note the warning, due to small cell sizes.

Example 22.2 may be calculated using the following script:

```

> x = c(652,1537,598,242,36,46,38,21,218,327,106,67)
> xmat = matrix(x, nrow=3, byrow=T)
> chisq.test(xmat,correct=T)

Pearson's Chi-squared test

data: xmat
X-squared = 51.6556, df = 6, p-value = 2.187e-09

```

## 22.6 Assumptions

To use the technique introduced here we assume that the sample selected is a true random sample. The  $\chi^2$  distribution for  $X^2$  is an approximation which is useful provided that the count in each cell is large enough. Many statistical computer packages will warn the user if any cell contains a count less than 5, which is a commonly used rule of thumb. Modifications of the technique presented here exist which are designed to correct for small cell counts. Yates' correction is the most common, and many computer packages give this as an option.

# Chapter 23

## ANOVA

We often have situations in which we have  $k$  random samples from  $k$  distinct populations with population means  $\mu_1, \dots, \mu_k$ . Interest is then in testing the hypothesis

$$\begin{aligned} H_o : \mu_1 &= \mu_2 = \dots = \mu_k \\ H_a : \mu_i &\neq \mu_j \text{ for some } i, j \end{aligned}$$

In other words, are there some differences between the means ( $H_a$ ) or are they all the same ( $H_o$ ).

### 23.1 Methodology

The technique we use is referred to as *analysis of variance*, or ANOVA. The data then has the following structure

Pop'n	Pop'n Mean	Sample Size	Sample	Sample Mean	Sum of Squares
1	$\mu_1$	$n_1$	$y_{11}, y_{12}, \dots, y_{1n_1}$	$\bar{y}_1$	$\sum_{i=1}^{n_1} (y_{1i} - \bar{y}_1)^2$
2	$\mu_2$	$n_2$	$y_{21}, y_{22}, \dots, y_{2n_2}$	$\bar{y}_2$	$\sum_{i=1}^{n_2} (y_{2i} - \bar{y}_2)^2$
:	:	:	:	:	:
$k$	$\mu_k$	$n_k$	$y_{k1}, y_{k2}, \dots, y_{kn_k}$	$\bar{y}_k$	$\sum_{i=1}^{n_k} (y_{ki} - \bar{y}_k)^2$

The groups may be referred to as *treatments*. Sometimes it is convenient to refer to a treatment as a *factor* or *factor variable*. The observations  $y_{ij}$  are then *responses*, and  $\mu_i$  is a *mean response*. Here, we only have one factor, so the procedure is referred to as *one-way ANOVA*. If the sample sizes  $n_i$  are equal, we may refer to a *balanced design*.

We also have the *total mean*

$$\begin{aligned}\bar{y} &= \frac{\text{sum of all observations}}{n_1 + n_2 + \dots + n_k} \\ &= \frac{n_1\bar{y}_1 + n_2\bar{y}_2 + \dots + n_k\bar{y}_k}{n_1 + n_2 + \dots + n_k}\end{aligned}$$

In order to develop a test statistic for the hypothesis, we define the *treatment sum of squares*

$$\text{SST} = \sum_{i=1}^k n_i(\bar{y}_i - \bar{y})^2$$

and the *error sum of squares*

$$\text{SSE} = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2.$$

It can be shown that if define the *total sum of squares* to be

$$\text{SSTO} = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2.$$

then

$$\text{SSTO} = \text{SST} + \text{SSE}.$$

The test statistic we use is then

$$F_{obs} = \frac{\text{SST}/(k-1)}{\text{SSE}/(n-k)}$$

where

$$n = n_1 + n_2 + \dots + n_k.$$

Given the form of SST we can see that if there are large differences among the sample means,  $F_{obs}$  will tend to be larger. To reject the null hypothesis we use the observed significance level defined by

$$\alpha_{obs} = P(F_{k-1,n-k} > F_{obs})$$

where  $F_{\nu_1, \nu_2}$  is a random variable with an *F distribution* with  $\nu_1$  *numerator degrees of freedom* and  $\nu_2$  *denominator degrees of freedom*. Most statistical software packages will calculate this significance level.

**Example 23.1** This example is due to Johnson and Bhattacharya (*Statistics: Principles and Methods*, Wiley, 3rd edition).

In an effort to improve the quality of recording tapes, the effects of four kinds of coatings A, B, C and D on reproduction quality are assessed by applying each to a separate sample of tape and measuring the resulting distortion. The results are given in the following table.

Coating	Sample	Sample Mean	Sum of Squares
A	10, 15, 8, 12, 15	$\bar{y}_1 = 12$	$\sum_{i=1}^5 (y_{1i} - \bar{y}_1)^2 = 38$
B	14, 18, 21, 15	$\bar{y}_2 = 17$	$\sum_{i=1}^4 (y_{2i} - \bar{y}_2)^2 = 30$
C	17, 16, 14, 15, 17, 15, 18	$\bar{y}_3 = 16$	$\sum_{i=1}^7 (y_{3i} - \bar{y}_3)^2 = 12$
D	12, 15, 17, 15, 16, 15	$\bar{y}_4 = 15$	$\sum_{i=1}^6 (y_{4i} - \bar{y}_4)^2 = 14$

We therefore have

$$\begin{aligned}
k &= 4 \\
n &= n_1 + n_2 + n_3 + n_4 \\
&= 5 + 4 + 7 + 6 \\
&= 22 \\
\hat{y} &= \frac{n_1 \bar{y}_1 + n_2 \bar{y}_2 + n_3 \bar{y}_3 + n_4 \bar{y}_4}{n_1 + n_2 + \dots + n_k} \\
&= \frac{5 \times 12 + 4 \times 17 + 7 \times 16 + 6 \times 15}{22} \\
&= 15.
\end{aligned}$$

The sums of squares are given by

$$\begin{aligned}
SSE &= \sum_{i=1}^5 (y_{1i} - \bar{y}_1)^2 + \sum_{i=1}^4 (y_{2i} - \bar{y}_2)^2 + \sum_{i=1}^7 (y_{3i} - \bar{y}_3)^2 + \sum_{i=1}^6 (y_{4i} - \bar{y}_4)^2 \\
&= 38 + 30 + 12 + 14 \\
&= 94
\end{aligned}$$

and

$$\begin{aligned}
SST &= n_1(\bar{y}_1 - \bar{y})^2 + n_2(\bar{y}_2 - \bar{y})^2 + n_3(\bar{y}_3 - \bar{y})^2 + n_4(\bar{y}_4 - \bar{y})^2 \\
&= 5(-3)^2 + 4(2)^2 + 7(1)^2 + 6(0)^2 \\
&= 68
\end{aligned}$$

giving test statistic

$$\begin{aligned}
F_{obs} &= \frac{SST/(k-1)}{SSE/(n-k)} \\
&= \frac{68/3}{94/18} \\
&= 4.34
\end{aligned}$$

The observed significance level can be calculated using the appropriate table or with a computer program, and can be found to be

$$\alpha_{obs} = .018$$

meaning that there is evidence that the four means are not identical. ■

## 23.2 ANOVA Table

The results of an ANOVA calculation are usually summarized in an *ANOVA summary table* of the following form

Source	SS	df	MS	
Between Treatment (or Treatment)	SST	$k - 1$	$MST = \frac{SST}{k-1}$	$F = \frac{MST}{MSE}$
Within Treatment (or Error)	SSE	$n - k$	$MSE = \frac{SSE}{n-k}$	
Total	SSTO	$n - 1$		

Where

$k$	Number of groups	
$n_i$	Sample size of group $i$	
$n$	Total sample size	$n_1 + \dots + n_k$
$\bar{y}_i$	Sample mean of group $i$	
$\bar{y}$	Total sample mean	$\frac{n_1\bar{y}_1 + \dots + n_k\bar{y}_k}{n_1 + \dots + n_k}$
SSE	Error sum of squares or Within Treatment SS	$\sum_{i=1}^k (n_k - 1)s_k^2$
SST	Treatment sum of Squares or Between Treatment SS	$\sum_{i=1}^k n_k(\bar{y}_k - \bar{y})^2$
SSTO	Total sum of squares	$SST + SSE$
MSE	Mean error sum of squares or Mean within Treatment SS	$\frac{SSE}{n-k}$
MST	Mean treatment sum of squares or Mean between Treatment SS	$\frac{SST}{k-1}$
$F$	F-ratio	$\frac{MST}{MSE}$

For the tape coating problem

$$\begin{aligned}
 n &= 22 \\
 k &= 4 \\
 SSE &= 94 \\
 MSE &= 94/(n - k) \\
 &= 5.22 \\
 SST &= 68 \\
 MST &= 68/(k - 1) \\
 &= 22.67 \\
 F &= 4.34
 \end{aligned}$$

The ANOVA table for this problem is then

Source	SS	df	MS	
Between Treatment	68	3	22.67	4.34
Within Treatment	94	18	5.22	
Total	162	21		

Recall that it is assumed that each sample comes from a population with possibly differing means, but with one common variance  $\sigma^2$ . It can be shown that  $MSE$  is an estimator of  $\sigma^2$ . In fact, if there are  $k = 2$  groups then the  $MSE$  is identical to the pooled sample variance  $S_p^2$  and plays the same role when  $k > 2$ .

### 23.3 Bonferroni Correction for Multiple Comparisons

We often encounter a situation in which we wish to report several confidence intervals or hypothesis tests. If we use a confidence level  $(1 - \alpha)$  for each confidence interval, or a significance level of  $\alpha$  for each hypothesis test, we must consider the fact that the probability of at least one error among all inference statements will be greater than  $\alpha$ .

A number of procedures exist with which to control *familywise error rate* (FWE), that is, the probability that among a set of  $m$  inference statements there is at least one error (the term *group* is sometimes used in place of ‘familywise’). The commonly used convention is that an error rate suitable for a single inference procedure should also be applied to multiple inferences, so that the FWE is commonly set to  $\alpha_{FWE} = 0.05$ . We can also refer to familywise (or group) confidence level  $1 - \alpha_{FWE}$ .

A large number of *multiple comparison* procedures exist, some specialized and others general. Probably the most commonly encountered method is known as the *Bonferroni correction procedure* (BCP), which is applicable, in principle, to any multiple comparison model. Recall Boole's inequality:

$$P(\cup_{i=1}^m E_i) \leq \sum_{i=1}^m P(E_i).$$

Suppose we are given  $m$  level  $(1 - \alpha)$  confidence intervals

$$E_i = \{ \text{the } i\text{th CI is incorrect} \}.$$

Then  $P(E_i) = \alpha$  and

$$P(\cup_{i=1}^m E_i) \leq m\alpha. \quad (23.1)$$

This means that all  $m$  confidence intervals are correct with a probability of at least  $1 - m\alpha$ . Therefore, in order to achieve a FWE of  $\alpha_{FWE}$ , we would need to use a confidence level of  $(1 - \alpha_{FWE}/m)$ .

**Example 23.2** Normally, to achieve a confidence level of 95% for a confidence interval

$$\bar{X} \pm z_{\alpha/2} \sigma / \sqrt{n} \quad (23.2)$$

we would set  $\alpha = 0.05$  and therefore use critical value  $z_{0.025} \approx 1.96$ . If we wanted to simultaneously report  $m = 4$  confidence intervals with FWE  $\alpha_{FWE} = 0.05$ , we would build separate level  $(1 - \alpha/m)$  confidence intervals. From (23.1), this would give

$$\alpha_{FWE} \leq m\alpha/m = \alpha,$$

so that it would be appropriate to set  $\alpha = \alpha_{FWE}$ . Therefore, in (23.2) we would use the critical value

$$z_{\alpha_{FWE}/(2m)} = z_{.05/8} = z_{.00625} \approx 2.5$$

for  $\alpha_{FWE} = 0.05$ . However, construction of the confidence intervals uses the same methodology once the Bonferroni correction has been applied.

The sample principle applies to hypothesis tests. If we want to report  $m$  hypothesis tests with a familywise Type I error of  $\alpha_{FWE}$  (that is, at least one Type I error among the  $m$  tests), then each test must be carried out with a significance level of  $\alpha_{FWE}/m$ . ■

## 23.4 Post hoc Analysis in ANOVA

If we conclude that there is some difference between means using the  $F$ -test, then we may wish to further explore how the means differ. A common way to achieve this is through the use of *pairwise multiple comparisons*.

Using the BCP we have

$$\bar{y}_i - \bar{y}_j \pm t_{\alpha/(m^2), n-k} \sqrt{MSE \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}$$

where  $m$  is the number of comparisons we wish to make (we need not be interested in all available comparisons).

If  $\mu_i$  and  $\mu_j$  are two group means, then a confidence interval for  $\mu_i - \mu_j$  is given by

$$\bar{y}_i - \bar{y}_j \pm q_\alpha \sqrt{\frac{MSE}{2} \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}$$

where  $q_\alpha$  is the critical value from the *studentized range* distribution with  $k-1$  treatment degrees of freedom and  $n-k$  error degrees of freedom (these have the same interpretation as the numerator and denominator degrees of freedom for the  $F$ -distribution). Tables for these critical values are available in most textbooks. This is known as *Tukey's pairwise procedure* or sometimes the *Tukey-Kramer pairwise procedure*. It should be noted that the procedure is approximate, unless the design is balanced.

Note that there will be  $k(k-1)/2$  comparisons. It needs to be stressed that the confidence level  $1-\alpha$  represents the probability that *all*  $k(k-1)/2$  confidence intervals are correct, and not just each one taken individually.

To continue with the tape coating problem, we have 18 error degrees of freedom and 3 treatment degrees of freedom so if we want a 95% confidence interval for all pairwise comparisons simultaneously we set

$$q_{0.05} = 4.00$$

and we also have

$$MSE = 5.22$$

with

Coating	$n_i$	Sample mean
A	$n_1 = 5$	$\bar{y}_1 = 12$
B	$n_2 = 4$	$\bar{y}_2 = 17$
C	$n_3 = 7$	$\bar{y}_3 = 16$
D	$n_4 = 6$	$\bar{y}_4 = 15$

giving pairwise confidence intervals

Pair	Confidence Interval
$\mu_1 - \mu_2:$	$-5 \pm 4.33$
$\mu_1 - \mu_3:$	$-4 \pm 3.78$
$\mu_1 - \mu_4:$	$-3 \pm 3.91$
$\mu_2 - \mu_3:$	$1 \pm 4.06$
$\mu_2 - \mu_4:$	$2 \pm 4.17$
$\mu_3 - \mu_4:$	$1 \pm 3.59$

We may conclude that  $\mu_1$  is significantly different from  $\mu_2$  and  $\mu_3$  but can make no other conclusions based on this procedure.

Note that there are many pairwise procedures, most notably the *Scheffe test* which is very conservative. This provides a FWE of  $\alpha_{FWE}$  of confidence intervals for all *contrasts*

$$C = \sum_{i=1}^n c_i \mu_i, \text{ where } \sum_{i=1}^k c_i = 0.$$

A pairwise comparison of the form  $\mu_i - \mu_j$  is a contrast of the form  $c_i = 1$ ,  $c_j = -1$ ,  $c_k = 0$  for  $k \neq i, j$ .

## 23.5 Nonparametric ANOVA

Recall that ANOVA may be thought of as an extension of the two-sample t-test for differences in mean to a  $K$ -sample test for differences in means, under the assumptions that variances are equal. Similarly, the *Kruskal-Wallis test* is an extension of the Wilcoxon rank sum test to  $K$  samples, and may be considered a nonparametric alternative to ANOVA. Under the null hypothesis,  $K$  samples are taken from  $K$  identical distributions (not necessarily normally distributed). We won't discuss the details of this test, but will note that the Kruskal-Wallis test is implemented in most statistical software packages. It would be appropriate to use whenever ANOVA might be used, but does not assume that the data is normally distributed.

## 23.6 Assumptions

The essential assumptions made for ANOVA are that population  $i$  has a  $N(\mu_i, \sigma^2)$  distribution. The means may differ between populations but variances do not. In addition, each sample is a true random sample, and the samples are independent of each other.

Of course, ANOVA is a technique which has received a great deal of attention by statistical practitioners, so that there are a wide variety of techniques which may be used when these assumptions do not hold.

# Chapter 24

## ANOVA in R

To fit an ANOVA model in R, we express the data as such a model. The variable  $Y$  is a single vector which contains all the variable.  $X$  is a single vector of *factors* which define the treatments. For example, to set up an ANOVA model in R, we can use the commands:

```
> y1 = rnorm(5,mean=10,sd=2.4)
> y2 = rnorm(6,mean=20,sd=2.4)
> y3 = rnorm(4,mean=20,sd=2.4)
>
> y = c(y1, y2, y3)
> x = c(rep(1,5), rep(2,6), rep(3, 4))
> x = as.factor(x)
> cbind(x,y)
      x         y
[1,] 1  6.526123
[2,] 1 10.639951
[3,] 1  8.128591
[4,] 1 10.922928
[5,] 1 13.934701
[6,] 2 20.502000
[7,] 2 22.109955
[8,] 2 22.575551
[9,] 2 23.177065
[10,] 2 19.509436
[11,] 2 19.890914
[12,] 3 15.414790
[13,] 3 18.572681
[14,] 3 20.668094
[15,] 3 15.777068
>
```

This simulates an ANOVA model with  $k = 3$  treatments, identified in the factor variable  $x$ . The sample sizes are  $n_1 = 5$ ,  $n_2 = 6$ ,  $n_3 = 4$ , with means  $\mu_1 = 10$ ,  $\mu_2 = \mu_3 = 20$ . The common variance is  $\sigma^2 = 2.4^2$ .

It is usually a good idea to plot the data, and this can be done using boxplots. The R model notation can be used to separate the groups:

```
> boxplot(y ~ x)
```

The plot is shown in Figure 24.1.

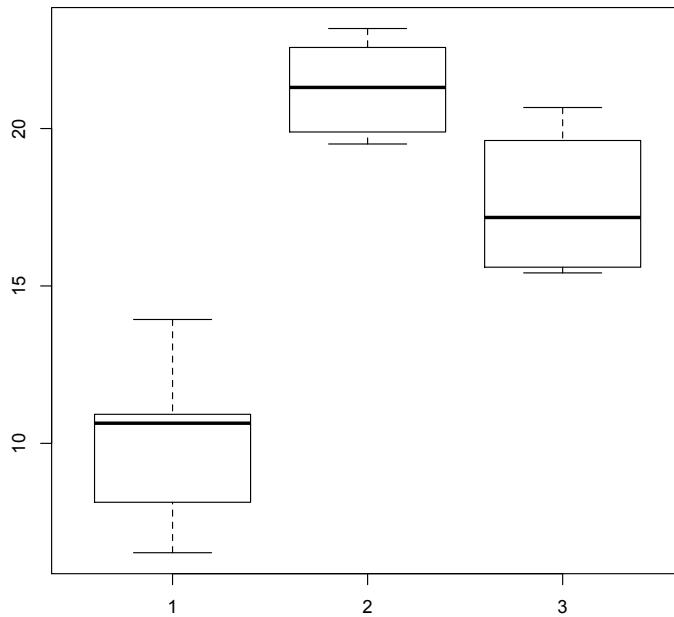


Figure 24.1: Multiple boxplots for ANOVA example

There are several ways to fit an ANOVA model in R. One dedicated function is `aov()` used as follows:

```
> fit = aov(y ~ x)
> summary(fit)

Df Sum Sq Mean Sq F value    Pr(>F)
x          2  352.0   176.0   33.85 1.17e-05 ***
Residuals 12   62.4     5.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that the fit itself can be stored as an object, which is generally good practice. Fit objects can then be used as input for generic functions, which provide summaries for the appropriate type of object. For example `summary()` gives for an `aov()` object the standard ANOVA table.

Tukey's pairwise procedure is also available for an ANOVA fit using the `TukeyHSD()` function (HSD refers to 'Honest Significant Difference'):

```
> fit.Tukey = TukeyHSD(fit)
> fit.Tukey
  Tukey multiple comparisons of means
  95% family-wise confidence level

Fit: aov(formula = y ~ x)

$x
    diff      lwr      upr      p adj
2-1 11.263695  7.579838 14.9475516 0.0000085
3-1  7.577699  3.496636 11.6587623 0.0009009
3-2 -3.685995 -7.613000  0.2410094 0.0665423

>
```

Notice that a new R object was produced by the `TukeyHSD()` function. If we use the generic `plot()` function we get the following plot (Figure 24.2):

```
> plot(fit.Tukey)
```

## 24.1 Equality of Variances

We have seen how to test for the equality of two variances. *Bartlett's Test* is a generalization to  $k$  variances suitable for ANOVA. This is available using the `bartlett.test()`, using the same model notation. This may use the same model notation:

```
> bartlett.test(y ~ x)

Bartlett test of homogeneity of variances

data: y by x
Bartlett's K-squared = 1.5712, df = 2, p-value = 0.4558
```

The large  $p$ -value means that the hypothesis of equality of variances (also known as *homoscedasticity*, as opposed to *heteroscedasticity*) was not rejected.

It is also possible to use `bartlett.test()` by input a list of samples:

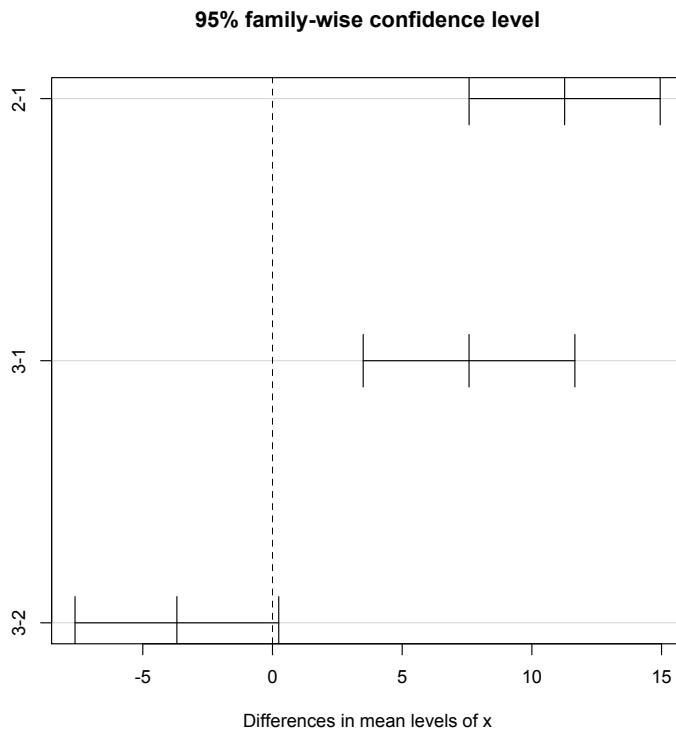


Figure 24.2: Graphical representation of Tukey's pairwise procedure.

```
> y.list = list(y1, y2, y3)
> y.list
[[1]]
[1] 6.526123 10.639951 8.128591 10.922928 13.934701

[[2]]
[1] 20.50200 22.10996 22.57555 23.17706 19.50944 19.89091

[[3]]
[1] 15.41479 18.57268 20.66809 15.77707

> bartlett.test(y.list)

Bartlett test of homogeneity of variances

data: y.list
Bartlett's K-squared = 1.5712, df = 2, p-value = 0.4558
```

&gt;

However, `aov()` cannot be used this way.

We finally note that a model can be converted to a list using the `split()` function:

```
> split(y,x)
$`1`
[1] 6.526123 10.639951 8.128591 10.922928 13.934701

$`2`
[1] 20.50200 22.10996 22.57555 23.17706 19.50944 19.89091

$`3`
[1] 15.41479 18.57268 20.66809 15.77707
```

&gt;

## 24.2 The Kruskal-Wallis Test for Nonparametric ANOVA

We have briefly introduced the Kruskal-Wallis test as an extention of the rank sum procedure to more than 2 samples. This is implemented in R using the `kruskal.test()` function, which is similar to `aov()`. For example, consider the simulated data:

```
> y1 = rnorm(5,mean=10,sd=2.4)
> y2 = rnorm(6,mean=20,sd=2.4)
> y3 = rnorm(4,mean=20,sd=2.4)
>
> y = c(y1, y2, y3)
> x = c(rep(1,5), rep(2,6), rep(3, 4))
> x = as.factor(x)
> cbind(x,y)
   x      y
[1,] 1 11.695035
[2,] 1  7.967687
[3,] 1  8.891760
[4,] 1 12.476237
[5,] 1 10.996793
[6,] 2 21.324256
[7,] 2 19.594350
[8,] 2 15.141688
```

```
[9,] 2 21.956811  
[10,] 2 19.251928  
[11,] 2 16.064112  
[12,] 3 21.472492  
[13,] 3 20.310484  
[14,] 3 26.817237  
[15,] 3 20.906802  
>  
> boxplot(y ~ x)  
>
```

The boxplot is shown in Figure 24.3. The data may be input into the

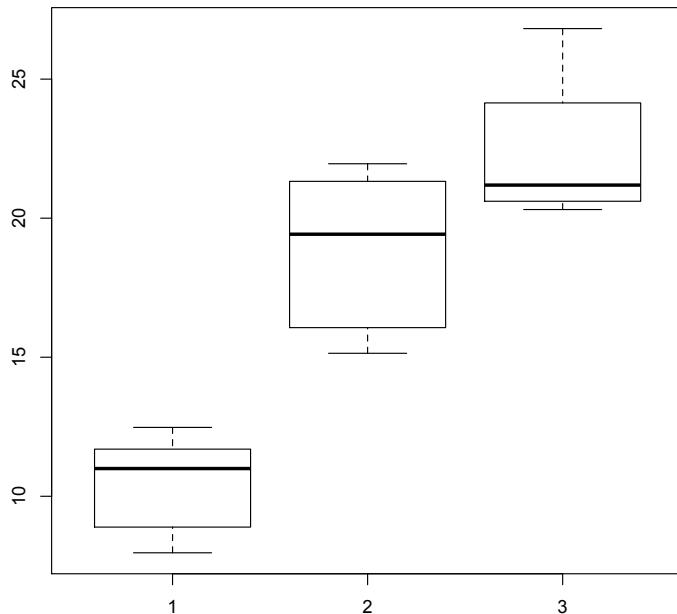


Figure 24.3: Multiple boxplots for Kruskal-Wallis test.

`kruskal.test()` function as a model:

```
> fit = kruskal.test(y ~ x)  
> summary(fit)  
Length Class Mode  
statistic 1 -none- numeric
```

```

parameter 1      -none- numeric
p.value   1      -none- numeric
method    1      -none- character
data.name 1      -none- character
> fit

Kruskal-Wallis rank sum test

data: y by x
Kruskal-Wallis chi-squared = 10.3958, df = 2, p-value = 0.005528

> stat = fit$statistic
> df = fit$parameter
> 1-pchisq(stat,df)
Kruskal-Wallis chi-squared
0.005528069
>

```

However, in this case the `summary()` function only lists the labels which define the list elements of the fit object. These labels provide access to the output quantities. For example, the significance level is calculated from a  $\chi^2$  statistic with 2 degrees of freedom. We can access the statistic and the degrees of freedom by references to `fit$statistic` and `fit$parameter`. We have justly illustrated this by recalculating the *p*-value from the output.

The function `kruskal.test()` also accepts multiple samples in list form. In addition, it has a `g` option which defines groups, permitting the data to be entered as a single array"

```

> y.list = list(y1, y2, y3)
> y.list
[[1]]
[1] 11.695035 7.967687 8.891760 12.476237 10.996793

[[2]]
[1] 21.32426 19.59435 15.14169 21.95681 19.25193 16.06411

[[3]]
[1] 21.47249 20.31048 26.81724 20.90680

> kruskal.test(y.list)

Kruskal-Wallis rank sum test

```

```
data: y.list
Kruskal-Wallis chi-squared = 10.3958, df = 2, p-value = 0.005528

> kruskal.test(y, g = x)

Kruskal-Wallis rank sum test

data: y and x
Kruskal-Wallis chi-squared = 10.3958, df = 2, p-value = 0.005528

>
```

## Chapter 25

# Linear Regression I

Consider the scatter plot in Figure 25.1 representing 392 automobiles. The horizontal axis gives the engine displacement in cubic inches and the vertical axis gives horsepower.

There seems to be a definite increasing trend in horsepower as engine displacement increases. If we set

$$\begin{aligned} X &= \text{Engine Displacement} \\ Y &= \text{Horsepower} \end{aligned}$$

then we should have approximately a linear relationship

$$Y = \beta_0 + \beta_1 X$$

where  $\beta_0$  and  $\beta_1$  are *coefficients* to be determined from the data. Looking at the scatter plot we see that the relationship will not be exact, so we introduce a random term  $\epsilon$  (*epsilon*) into the equation.

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

The usual terminology is to refer to  $Y$  as the *dependent variable* and to  $X$  as the *independent variable* or *predictor*. Here, there is only one predictor  $X$ , so the model is termed *simple linear regression*. When more predictors are used, for example  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ , the model is termed *multiple linear regression*.

If we have  $n$  pairs of dependent and independent observations

$$(X_1, Y_1), \dots, (X_n, Y_n)$$

then the regression equation can be written in terms of the sample

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \dots, n. \quad (25.1)$$

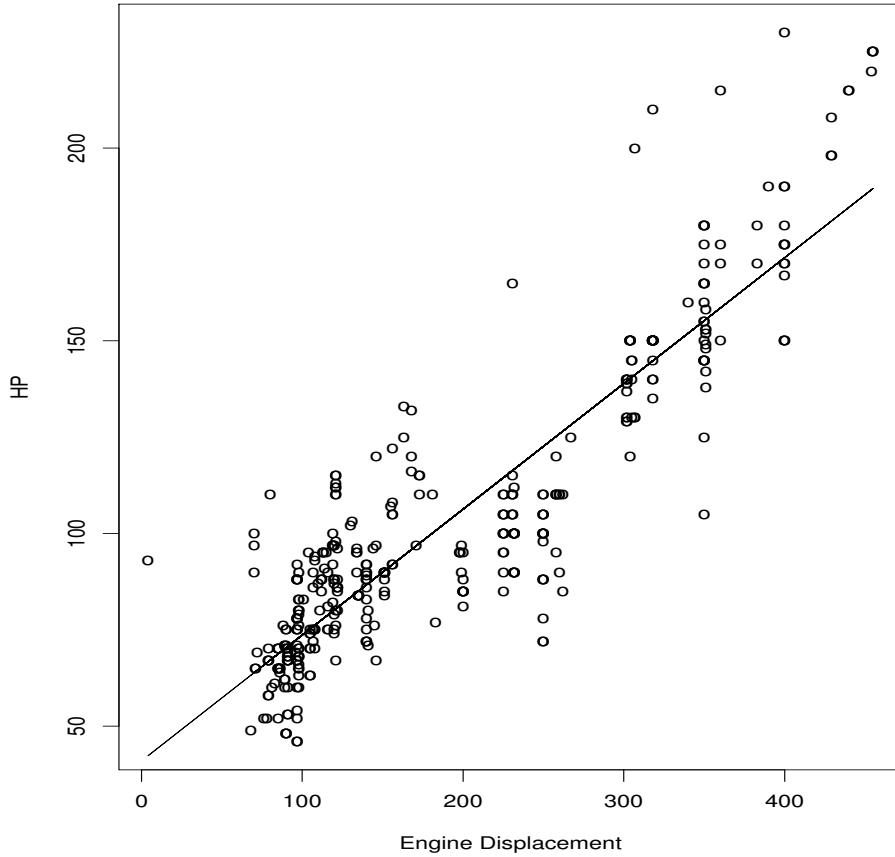


Figure 25.1: Scatter plot of automobile data

Here, we assume that the *error terms*  $\epsilon_1, \dots, \epsilon_n$  form a random sample from  $N(0, \sigma^2)$ . We do not observe the error terms (unlike  $X_i$  and  $Y_i$ ), but we can estimate  $\sigma^2$ .

The *linear least squares coefficients* are given by

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X}.\end{aligned}$$

With a large enough sample size we have estimates

$$\hat{\beta}_0 \approx \beta_0 \text{ and } \hat{\beta}_1 \approx \beta_1,$$

giving the estimated relationship between  $X$  and  $Y$

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X.$$

We also have the *predicted responses*

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i.$$

for each sample pair  $i = 1, \dots, n$ . Of course, we may construct a predicted response for a predictor value not represented in the sample, that is,

$$\hat{Y}_x = \hat{\beta}_0 + \hat{\beta}_1 x$$

is the predicted response for a predictor value  $X = x$ .

In the above example we have

$$\begin{aligned}\hat{\beta}_1 &= 0.327 \\ \hat{\beta}_0 &= 41.002,\end{aligned}$$

and this line is drawn in Figure 25.1.

Most statistical software implements linear regression, giving output in the following format

Model	Unstandardized Coefficients				
	B	Std. Error	t	Sig.	
1	(Constant) 41.002	1.792	22.884	.000	
	Engine Displacement (cu. inches) .327	.008	40.500	.000	

Least squares coefficients can be taken directly from this table. If we wish to construct a level  $(1 - \alpha)100\%$  confidence interval for  $\beta_0$  and  $\beta_1$  we may use

$$\begin{aligned}\hat{\beta}_0 &\pm t_{n-2,\alpha/2} \times \text{Std. Error for } \hat{\beta}_0 \\ \hat{\beta}_1 &\pm t_{n-2,\alpha/2} \times \text{Std. Error for } \hat{\beta}_1\end{aligned}$$

where the standard error may be taken from the table. Note that the appropriate degrees of freedom for the  $t$ -distribution critical values are  $n - 2$ . If  $n$  is very large, we may use the standard normal critical value  $z_{\alpha/2}$  instead. For the above example we have 95% confidence intervals

$$\begin{aligned}CI_{.95} &= 41.002 \pm 1.96 \times 1.792 \\ &= 41.002 \pm 3.5\end{aligned}$$

for  $\beta_0$  and

$$\begin{aligned} CI_{.95} &= 0.327 \pm 1.96 \times 0.008 \\ &= 0.327 \pm 0.0157 \end{aligned}$$

for  $\beta_1$ .

An important hypothesis test is

$$\begin{aligned} H_o : \beta_1 &= 0 \\ H_a : \beta_1 &\neq 0. \end{aligned}$$

This tells us whether or not there is any relationship between the dependent and independent variable. The observed significance level can be read directly from the table in the last column. Here, the observed significance level is given as 0, which means that there is strong evidence of a linear relationship between engine displacement and horsepower.

## 25.1 Residuals

The basic assumption used here is that the error terms  $\epsilon_1, \epsilon_2, \dots, \epsilon_n$  where

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

are equivalent to a random sample from a normal distribution with mean 0 and variance  $\sigma^2$ . Implicit in this formulation is the assumption that there is a linear relationship between  $X$  and  $Y$ .

Of course, the  $\epsilon_i$ 's cannot be directly observed, but they can be estimated by the *residuals*, given by

$$e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i = Y_i - \hat{Y}_i$$

once the regression has been calculated. There are several ways to use the residuals to check the assumptions.

1. Draw a scatter plot of the points  $(e_i, \hat{Y}_i)$  where  $\hat{Y}_i$  is the *predicted value*

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i.$$

If the assumptions are satisfied there should be no pattern.

- (a) Check to see if the variation of the residuals appears to increase or decrease systematically. If so, this means that the variance of the error terms is not constant.

- (b) If large groups of residuals located next to each other appear to be all above or all below zero, then the assumption that the error terms are independent of each other may be incorrect. This is a frequent occurrence when the  $X_i$ 's represent sequential points in time.
  - (c) If the residuals appear to suggest some functional form, then the assumption of a linear relationship between  $X$  and  $Y$  may be incorrect.
2. To check for the assumption of normality of the error terms, construct a normal probability plot of the residuals. Departures from linearity indicate departures from normality of the error terms.

As a final remark linear regression, like ANOVA, is a widely used tool, and many techniques exist which may be used when some of these assumptions are not valid.

**Example 25.1** To continue with the automobile section we present a residual plot and a normal probability plot (Figure 25.2).

The residual plot shows a somewhat different behavior below and above 100 horsepower, which might be investigated. Other than that, no systematic departure from the assumption of no pattern is indicated.

The normal probability plot is approximately linear, except for the two extreme regions. This indicates that the normal distribution might not be accurate for very small tail probabilities, but otherwise should suffice as an approximation. ■

**Example 25.2** As a second example, we reexamine the scatter plot presented in Example 4.3. The scatter plot is presented with a linear regression fit (Figure 25.3). From the scatter plot, we can see that although there is a strong relationship between miles per gallon and horsepower, it is not a strictly linear one. Accordingly, the residual plot indicates a systematic functional form, suggesting that a linear fit is not the appropriate one. ■

## 25.2 ANOVA approach

In linear regression, variation may be decomposed in a manner similar to that of ANOVA. We start with the *error sum of squares*

$$SSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2,$$

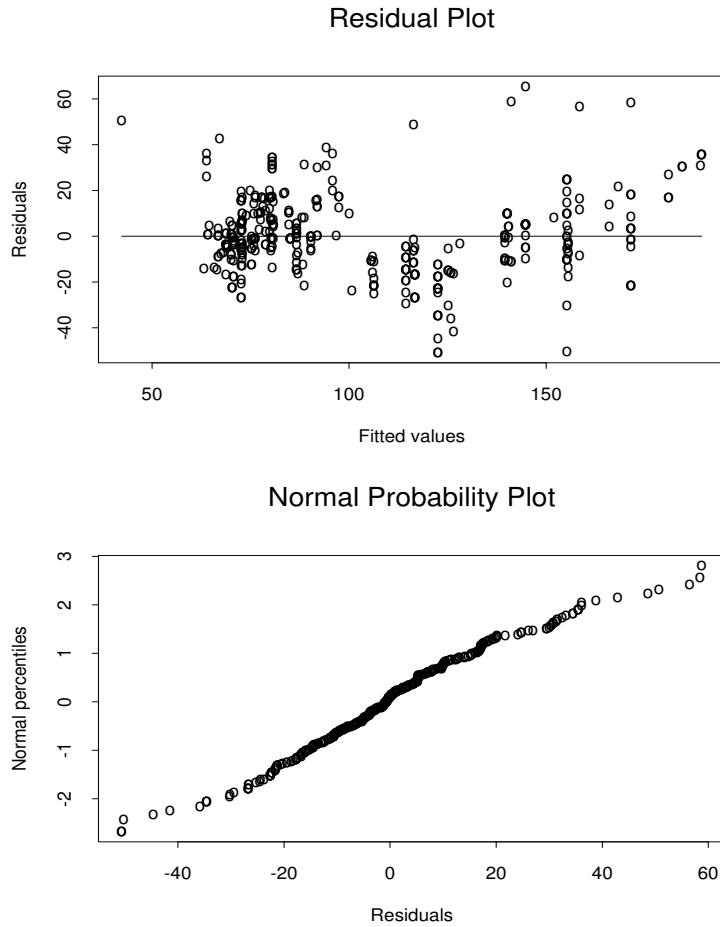


Figure 25.2: Residual plot and normal probability plot of residuals for Example 25.1

and the *mean error sum of squares*

$$MSE = \frac{SSE}{n - 2}.$$

The  $MSE$  is analogous to the MSE encountered in ANOVA, with  $K = 2$  treatments corresponding to the 2 unknown parameters  $\beta_0$  and  $\beta_1$ . In fact, the MSE functions as an estimate of the variance  $\sigma^2$  encountered in the distribution  $\epsilon_i \sim N(0, \sigma^2)$ :

$$\hat{\sigma}^2 = MSE \approx \sigma^2.$$

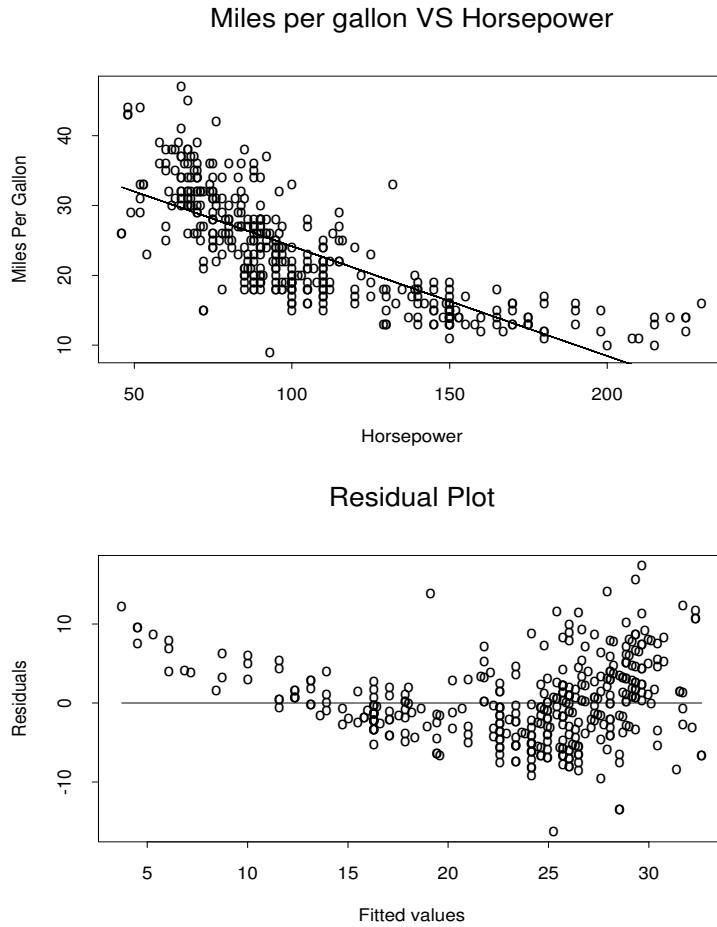


Figure 25.3: Scatter plot of MPG vs. Horsepower with linear regression fit, and residual plot for Example 25.2

We then have, as for ANOVA, the total sum of squares

$$SSTO = \sum_{i=1}^n (Y_i - \bar{Y})^2.$$

By convention, instead of the *treatment sum of squares*  $SST$  we define the *regression sum of squares*

$$SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2,$$

noting that the two quantities serve similar functions. It can be shown that

$$SSTO = SSR + SSE.$$

This means that, as in ANOVA, the total variation  $SSTO$  can be decomposed into variation  $SSR$  explained by the model and variation  $SSE$  attributable to the error terms  $\epsilon_i$ . The ANOVA table for simple linear regression therefore looks like:

Source	SS	df	MS	
Regression	$SSR$	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
Error	$SSE$	$n - 2$	$MSE = \frac{SSE}{n-2}$	
Total	$SSTO$	$n - 1$		

As in ANOVA,  $F$  has an  $F$ -distribution with 1 numerator and  $n - 2$  denominator degrees of freedom under the hypothesis

$$H_0 : \beta_1 = 0.$$

A quantity of considerable importance is the *coefficient of determination*

$$R^2 = 1 - \frac{SSE}{SSTO} = \frac{SSR}{SSTO}$$

which can be interpreted as the proportion of the total variation explainable by the model. We always have  $0 \leq R^2 \leq 1$ , so that larger values (say,  $R^2 \geq 0.25$ ) mean that the predictor  $X$  has significant explanatory power.

### 25.3 The Relationship Between Linear Regression and Correlation

It is important to note the similarity between the definition of  $r$  and the estimate of the slope  $\beta_1$  for simple linear regression:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

This means  $r$  and  $\hat{\beta}_1$  have a close relationship:

$$\begin{aligned} r &= r \frac{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}}{\sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \\ &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \times \frac{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}}{\sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \\ &= \hat{\beta}_1 \sqrt{\frac{S_X^2}{S_Y^2}} \end{aligned} \tag{25.2}$$

where  $S_X^2$  and  $S_Y^2$  are the samples variances of the  $X_i$ 's and  $Y_i$ 's.

When deducing the distribution properties of  $r$ , it is usually assumed that  $X$  and  $Y$  together possess a *bivariate normal distribution*. This means that  $X$  and  $Y$  are both normally distributed, and also possess a linear relationship of the form

$$Y = \beta_0 + \beta_1 X + \epsilon \quad (25.3)$$

where  $\beta_0$  and  $\beta_1$  are constants and  $\epsilon \sim N(0, \sigma^2)$  is independent of both  $X$  and  $Y$ . It can be shown that when (25.3) holds the correlation between  $X$  and  $Y$  is

$$\rho = \rho_{XY} = \beta_1 \frac{\sigma_X}{\sigma_Y},$$

which is directly comparable to (25.2) (when convenient, subscripts may be added to the symbols  $r$  or  $\rho$  to identify the relevant random variables). Of course, one important difference remains between (25.3) and the simple linear regression model, namely that for simple linear regression  $X$  is interpreted as a nonrandom predictor variable, whereas in (25.3)  $X$  is a random variable. Nonetheless, both models depend on the very specific notion of linear dependence between two variables.

It is important to note that assuming only that  $X$  and  $Y$  are normally distributed does not suffice to define the bivariate normal distribution. The assumption of a linear relationship is also needed.

## 25.4 Assumptions

The assumptions underlying simple linear regression are all implied in the model defined in (25.1):

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \dots, n. \quad (25.4)$$

Essentially, we assume  $Y_i \sim N(\mu_i, \sigma^2)$  for some  $\sigma^2$  which does not vary with the index  $i$ , where the means  $\mu_i$  are given by

$$\mu_i = \beta_0 + \beta_1 X_i. \quad (25.5)$$

Finally, the responses  $Y_i$  are assumed to be independent. This is equivalent to assuming that  $\epsilon_1, \dots, \epsilon_n$  is a random sample from distribution  $N(0, \sigma^2)$ , and the response  $Y_i$  is given by (25.1).

## Chapter 26

# Linear Regression II

In this section we consider in more detail inference for linear regression. We will emphasize the simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \dots, n,$$

but the ideas can be generalized when additional predictors are added.

Usually, there is greater interest in  $\beta_1$ , the *slope* of the regression line, than  $\beta_0$ , the point on the vertical response axis intercepted by the regression line at predictor value  $X = 0$  (hence  $\beta_0$  is commonly known as the *intercept*). This is because the motivation for regression is usually to determine a relationship between the dependent and independent variables, and a relationship can be said to exist between them if and only if the slope  $\beta_1$  is not zero.

We may consider two estimation problems. The mean response for predictor value  $x$  is

$$\mu_x = \beta_0 + \beta_1 x.$$

In principle, we may consider  $\mu_x$  for any value  $x$ , even if  $x$  does not equal the value of any predictor in a given sample. However, it is usually not recommended that  $x$  be *extrapolated* beyond the range of the observed predictors. If we have some reason to set  $x > \max_i X_i$  or  $x < \min_i X_i$ , it should be noted in any report that the resulting inference represents an extrapolation beyond the observed range of the predictor variables. Of course, the intercept  $\beta_0$  is a special case of  $\mu_x$ , in particular,

$$\beta_0 = \beta_0 + \beta_1 \times 0 = \mu_0,$$

however,  $\beta_1$  cannot be expressed as  $\mu_x$  for some  $x$  in this way.

## 26.1 Inference of Regression Parameters

We may define a general parameter  $\beta_i$ , and note that it's inference assumes a general form (we have, so far, encountered  $\beta_0$  and  $\beta_1$  for simple linear regression). We have seen estimates from Section 18

$$\hat{\beta}_i \approx \beta_i, \quad i = 0, 1$$

and we may add

$$\hat{\mu}_x \approx \hat{\beta}_0 + \hat{\beta}_1 x.$$

Note that the *predicted responses*

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \quad i = 1, \dots, n,$$

are special cases  $\hat{Y}_i = \hat{\mu}_{X_i}$ , and are commonly known as *fitted values*, since the estimated regression line passes through the points  $(X_i, \hat{Y}_i)$ .

Under the assumptions given in Section 18.3, in particular, that the error terms  $\epsilon_1, \dots, \epsilon_n$  are an independent random sample from  $N(0, \sigma^2)$  for some fixed variance  $\sigma^2$ , we have

$$\hat{\beta}_i \sim N\left(\beta_i, \sigma_{\hat{\beta}_i}^2\right),$$

and

$$\hat{\mu}_x \sim N\left(\mu_x, \sigma_{\hat{\mu}_x}^2\right).$$

It is worth noting at this point that  $\hat{\beta}_i$  and  $\hat{\mu}_x$  are *unbiased* estimates of  $\beta_i$  and  $\mu_x$ , since

$$E[\hat{\beta}_i] = \beta_i \text{ and } E[\hat{\mu}_x] \sim \mu_x,$$

(not all commonly used estimators are unbiased).

For simple linear regression, the values of  $\sigma_{\hat{\beta}_i}^2$  and  $\sigma_{\hat{\mu}_x}^2$  are well-known:

$$\sigma_{\hat{\beta}_1}^2 = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \tag{26.1}$$

and

$$\sigma_{\hat{\mu}_x}^2 = \sigma^2 \left[ \frac{1}{n} + \frac{(x - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right] \tag{26.2}$$

where we have mean value of the predictor:

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}.$$

Since  $\beta_0 = \mu_0$  we can obtain directly from (26.2) the variance of  $\hat{\beta}_0$  by substituting  $x = 0$ :

$$\sigma_{\hat{\beta}_0}^2 = \sigma^2 \left[ \frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]. \tag{26.3}$$

As we might expect, the values of  $\sigma_{\hat{\beta}_i}^2$  and  $\sigma_{\hat{\mu}_x}^2$  directly depend on error variance  $\sigma^2$ , which is usually unknown. Of course, we have estimate from Section 18.2,

$$\hat{\sigma}^2 = MSE = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2} \approx \sigma^2,$$

so we replace  $\sigma^2$  in (26.1), (26.2) and (26.3) with  $\hat{\sigma}^2$ , to obtain the *standard errors*

$$S_{\hat{\beta}_1} = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}}, \quad (26.4)$$

$$S_{\hat{\mu}_x} = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}} \quad (26.5)$$

and

$$S_{\hat{\beta}_0} = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}} \quad (26.6)$$

### 26.1.1 Confidence Intervals for Simple Linear Regression

Given the standard errors for  $\beta_i$  or  $\mu_x$  a level  $(1 - \alpha)$  confidence interval for  $\beta_i$  is given by

$$\hat{\beta}_i \pm t_{n-2,\alpha/2} \times S_{\hat{\beta}_i},$$

or for  $\mu_x$  by

$$\hat{\mu}_x \pm t_{n-2,\alpha/2} \times S_{\hat{\mu}_x},$$

where  $t_{n-2,\alpha/2}$  is the  $\alpha/2$  critical value for a  $t$ -distribution with  $n - 2$  degrees of freedom.

### 26.1.2 Hypothesis Tests for Simple Linear Regression

If we wish to test against a hypothesis

$$H_o : \beta_i = \beta'_i \quad (26.7)$$

we use statistic

$$T = \frac{\hat{\beta}_i - \beta'}{S_{\hat{\beta}_i}}$$

which, under the hypothesis defined in Equation (26.7) has a  $t$ -distribution with  $n - 2$  degrees of freedom.

The most common hypothesis test in the context of simple linear regression is obtained by setting hypothetical value  $\beta'_1 = 0$ , that is, the two-sided test:

$$H_o : \beta_1 = 0 \text{ against } H_a : \beta_1 \neq 0,$$

which gives observed significance level

$$\alpha_{obs} = 2P(T \leq -|T_{obs}|)$$

where

$$T_{obs} = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}}$$

and  $T$  has a  $t$ -distribution with  $n-2$  degrees of freedom. When suitable, one-sided hypothesis tests can be carried out as discussed in previous chapters.

### 26.1.3 Prediction Intervals for Simple Linear Regression

Let's define a random variable

$$Y_x \sim N(\mu_x, \sigma^2),$$

which can be interpreted as a *future response* from a linear model

$$Y_x = \beta_0 + \beta_1 x + \epsilon$$

for a given predictor value  $X = x$ , and  $\epsilon \sim N(0, \sigma^2)$ . We might wish to place *prediction bounds* on  $Y_x$ , that is, values  $Y_L, Y_U$  for which

$$P(Y_L \leq Y_x \leq Y_U) = 1 - \alpha.$$

We might set  $1 - \alpha = 95\%$ . If  $\beta_0, \beta_1, \sigma^2$  are known, this is easy to do:

$$\begin{aligned} Y_L &= \mu_x - z_{\alpha/2} \sigma, \\ Y_U &= \mu_x + z_{\alpha/2} \sigma. \end{aligned}$$

Otherwise, we estimate  $\mu_x$  and  $\sigma^2$ , and the prediction interval can be based on the deviation

$$D = Y_x - \hat{\mu}_x,$$

that is, the deviation of a future response  $Y_x$  from its estimated mean  $\hat{\mu}_x$ . At this point we note that  $Y_x$ , being some future response, is independent of the data used to estimate  $\hat{\mu}_x$ . This means  $Y_x$  and  $\hat{\mu}_x$  are independent, so that the variance of  $D$  is

$$\begin{aligned} var(D) &= var(Y_x) + var(\hat{\mu}_x) \\ &= \sigma^2 + \sigma_{\hat{\mu}_x}^2 \\ &= \sigma^2 \left[ 1 + \frac{1}{n} + \frac{(x - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right] \end{aligned}$$

making use of Equation (26.2). This leads to level  $(1 - \alpha)$  prediction interval

$$\hat{\mu}_x \pm t_{n-2, \alpha/2} \times \hat{\sigma} \left[ 1 + \frac{1}{n} + \frac{(x - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]^{1/2}.$$

### 26.1.4 Calculations Based on Sums of Squares

Despite the apparent complexity of the computations associated with linear regression, they can be organized around the 5 quantities

$$\sum_{i=1}^n X_i, \quad \sum_{i=1}^n Y_i, \quad \sum_{i=1}^n X_i^2, \quad \sum_{i=1}^n Y_i^2, \quad \sum_{i=1}^n X_i Y_i$$

from which we derive quantities

$$\begin{aligned}\bar{X} &= \frac{\sum_{i=1}^n X_i}{n} \\ \bar{Y} &= \frac{\sum_{i=1}^n Y_i}{n} \\ SS_X &= \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - n^{-1} \left( \sum_{i=1}^n X_i \right)^2 \\ SS_Y &= \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n Y_i^2 - n^{-1} \left( \sum_{i=1}^n Y_i \right)^2 \\ SS_{XY} &= \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^n X_i Y_i - n^{-1} \left( \sum_{i=1}^n X_i \right) \left( \sum_{i=1}^n Y_i \right).\end{aligned}$$

These quantities appeared in Section 4 as  $\bar{X}_n, \bar{Y}_n, SS_X(n), SS_Y(n), SS_{XY}(n)$ , but we omit reference to sample size  $n$  here for convenience.

The relevant quantities then become

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{SS_{XY}}{SS_X}, \\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X}, \\ \hat{\mu}_x &= \hat{\beta}_0 + \hat{\beta}_1 x.\end{aligned}$$

This means estimates are most conveniently calculated in the order  $\hat{\beta}_1, \hat{\beta}_0$  and  $\hat{\mu}_x$  as required.

We next calculate  $SSE$  and  $SSTO$ , following which we may calculate any required standard errors. First

$$SSTO = SS_Y.$$

Although the calculation of  $SSE$  is not, at first, as straightforward, it can be shown that

$$SSE = \sum_{i=1}^n Y_i^2 - \hat{\beta}_0 \sum_{i=1}^n Y_i - \hat{\beta}_1 \sum_{i=1}^n X_i Y_i.$$

giving

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n - 2}$$

and coefficient of determination

$$R^2 = 1 - \frac{SSE}{SSTO}.$$

At this point we may calculate standard errors:

$$S_{\hat{\beta}_1} = \frac{\hat{\sigma}}{\sqrt{SS_X}},$$

$$S_{\hat{\mu}_x} = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x - \bar{X})^2}{SS_X}}$$

and

$$S_{\hat{\beta}_0} = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{SS_X}},$$

with  $(1 - \alpha)$  prediction interval for  $Y_x$

$$\hat{\mu}_x \pm t_{n-2,\alpha/2} \times \hat{\sigma} \left[ 1 + \frac{1}{n} + \frac{(x - \bar{X})^2}{SS_X} \right]^{1/2}.$$

**Example 26.1** The following example is due to Devore (Example 12.10) Devore (1995). Suppose we are given data ( $n = 11$ ):

X = 16.1 31.5 21.5 22.4 20.5 28.4 30.3 25.6 32.7 29.2 34.7  
Y = 4.41 6.81 5.26 5.99 5.92 6.14 6.84 5.87 7.03 6.89 7.87

We wish to construct a CI for  $\beta_1$ . We have the summary

$$\begin{aligned} \sum_{i=1}^n X_i &= 292.90 \\ \sum_{i=1}^n Y_i &= 69.03 \\ \sum_{i=1}^n X_i^2 &= 8141.75 \\ \sum_{i=1}^n Y_i^2 &= 442.1903 \\ \sum_{i=1}^n X_i Y_i &= 1890.200. \end{aligned}$$

We then have

$$\begin{aligned}\bar{X} &= 292.9/11 = 26.627 \\ \bar{Y} &= 69.03/11 = 6.275 \\ SS_X &= 8141.75 - (292.9^2)/11 = 342.622 \\ SS_Y &= 442.1903 - (69.03^2)/11 = 8.996 \\ SS_{XY} &= 1890.20 - 292.9 * 69.03/11 = 52.119,\end{aligned}$$

Giving coefficient estimates:

$$\begin{aligned}\hat{\beta}_1 &= \frac{SS_{XY}}{SS_X} = \frac{52.119}{342.622} = 0.152 \\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} = 6.275 - 0.1520 \times 26.627 = 2.228.\end{aligned}$$

The next step is to calculate  $SSE$ :

$$\begin{aligned}SSE &= \sum_{i=1}^n Y_i^2 - \hat{\beta}_0 \sum_{i=1}^n Y_i - \hat{\beta}_1 \sum_{i=1}^n X_i Y_i \\ &= 442.1903 - 2.228 \times 69.03 - 0.152 \times 1890.200 \\ &= 1.08,\end{aligned}$$

then

$$\hat{\sigma}^2 = SSE/(n - 2) = 1.08/9 = 0.12.$$

Then

$$S_{\hat{\beta}_1} = \frac{\hat{\sigma}}{\sqrt{SS_X}} = \frac{\sqrt{0.12}}{\sqrt{342.622}} = 0.019.$$

A level 95% confidence interval is then

$$0.152 \pm t_{9,\alpha/2} S_{\hat{\beta}_1} = 0.152 \pm 2.262 \times 0.019 = 0.152 \pm 0.042.$$

To calculate the coefficient of determination we set

$$SSTO = SS_Y = 8.992$$

so that

$$R^2 = 1 - SSE/SSTO = 1 - 1.08/8.992 = 0.88.$$

The high value for  $R^2$  is evident in Figure 26.1. ■

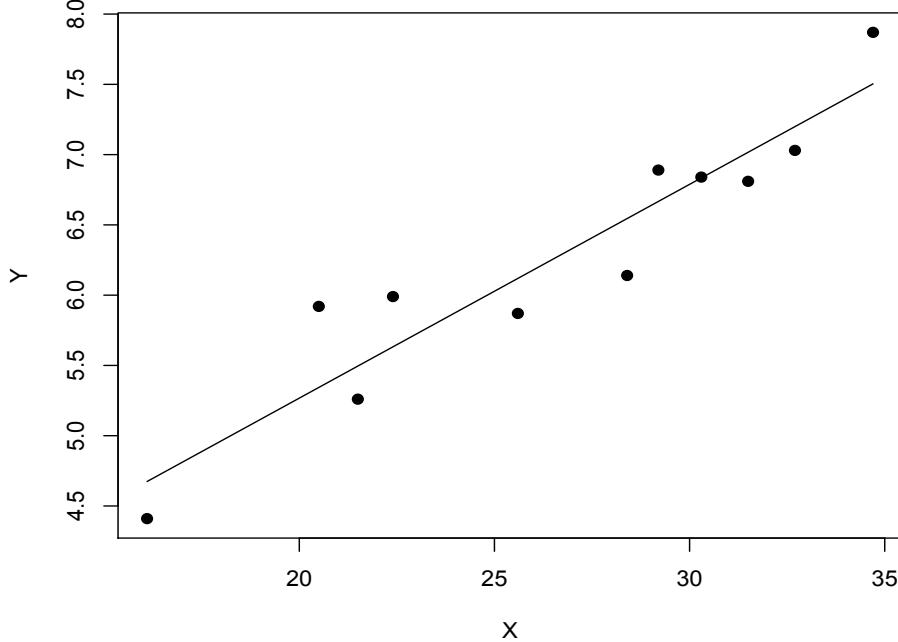


Figure 26.1: Scatter plot and regression fit for Example 26.1

## 26.2 Multiple Linear Regression

In contrast with simple regression, *multiple regression* permits  $q \geq 1$  predictors:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_q X_{qi} + \epsilon_i, \quad i = 1, \dots, n. \quad (26.8)$$

The predictors  $X_{ji}$  use a double subscript notation, where  $j$  refers to the predictor, and  $i$  refers to the sample. So, instead of observations in pairs  $(Y_i, X_i)$  for simple linear regression, observations come in the form  $(Y_i, X_{1i}, X_{2i}, \dots, X_{qi})$  for  $i = 1, \dots, n$ . In the context of multiple regression, it is usually the practice to refer to the  $j$ th predictor as  $X_j$ , on the understanding that a second subscript is needed to refer to the actual data. As in simple linear regression, the error terms  $\epsilon_1, \dots, \epsilon_n$  are a random sample from  $N(0, \sigma^2)$ , so that

$$Y_i \sim N(\mu_i, \sigma^2)$$

where

$$\mu_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_q X_{qi}, \quad i = 1, \dots, n$$

defines the *linear regression function*.

For each coefficient  $\beta_i$  we may obtain a *least squares estimate*  $\hat{\beta}_i$  and standard error  $S_{\hat{\beta}_i}$ . Their calculation requires techniques from matrix algebra which are beyond the scope of this course, so we rely on statistical computing. As in simple linear regression we have predicted, or fitted, values

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \dots + \hat{\beta}_q X_{qi}, \quad i = 1, \dots, n,$$

residuals

$$e_i = Y_i - \hat{Y}_i, \quad i = 1, \dots, n,$$

error sum of squares

$$SSE = \sum_{i=1}^n e_i^2,$$

and total sum of squares

$$SSTO = SS_Y = \sum_{i=1}^n (Y_i - \bar{Y})^2.$$

The regression sum of squares is similarly obtained from the equality

$$SSTO = SSR + SSE.$$

We also have the various mean sums of squares. The degrees of freedom associated with SSTO remains  $n - 1$ , but for SSE it is now  $n - (q + 1)$ , and for SSR it is  $q$ , giving

$$\begin{aligned} MSE &= \frac{SSE}{n - (q + 1)}, \\ MSR &= \frac{SSR}{q}. \end{aligned}$$

As in the simple linear regression case, an estimate of the error variance  $\sigma^2$  is given by

$$\sigma^2 \approx \hat{\sigma}^2 = MSE.$$

Confidence intervals for each coefficient  $\beta_j$  are given by

$$\hat{\beta}_j \pm t_{n-(q+1), \alpha/2} \times S_{\hat{\beta}_j}.$$

A test against null hypothesis

$$H_0 : \beta_j = \beta'_j$$

can be based on test statistic

$$T = \frac{\hat{\beta}_j - \beta'_j}{S_{\hat{\beta}_j}},$$

which under  $H_0$  has a  $t$ -distribution with  $n - (q + 1)$  degrees of freedom. If the null hypothesis  $H_0 : \beta_j = 0$  can be rejected, we may conclude that the response depends on the  $j$ th predictor  $X_j$  (in addition, possibly, to other predictors). Otherwise, there is no relationship between  $X_j$  and the response, and this predictor need not be included in the model (we usually include  $\beta_0$  in the model without the need of any formal inference).

### 26.2.1 ANOVA tables for multiple linear regression

The ANOVA table extends naturally to the multiple linear regression case:

Source	SS	df	MS	
Regression	SSR	$q$	$MSR = \frac{SSR}{q}$	$F = \frac{MSR}{MSE}$
Error	SSE	$n - (q + 1)$	$MSE = \frac{SSE}{n-(q+1)}$	
Total	SSTO	$n - 1$		

Here  $F$  has an  $F$ -distribution with  $q$  numerator and  $n - (q + 1)$  denominator degrees of freedom under the hypothesis

$$H_0 : \beta_i = 0, i = 1, \dots, q.$$

Note that this hypothesis does not specify that the intercept  $\beta_0$  is 0.

In the context of multiple regression the *coefficient of multiple determination* is

$$R^2 = 1 - \frac{SSE}{SSTO} = \frac{SSR}{SSTO}.$$

This definition is equivalent to the coefficient of determination defined for simple linear regression, but the alternative terminology emphasizes the influence on  $R^2$  of the number of parameters in the model.

### 26.2.2 Full and Reduced Models

Before we consider an actual example, it is important to understand the concept of the *full and reduced models*. Suppose we begin with the model (26.8) with  $q$  predictors. If any coefficient  $\beta_i$  is actually 0, there is no need to include it in the model. Of course, we don't know the exact value of  $\beta_i$ , but we might conclude on a statistical basis that it is not significantly different from 0, and so on that basis we can decide which predictors to keep in the model. It might seem that all we need to do is test each coefficient separately, keeping only those for which the null hypothesis  $H_0 : \beta_i = 0$  is rejected. There are two concerns with this approach. First, two predictors may be correlated with each other. When this happens, the respective coefficients may become difficult to interpret independently. In this case it is better to assess the predictive ability of the model as a whole. In addition, separate

inferences formally require multiple testing procedures, the application of which can be cumbersome when used to select predictors for inclusion.

One basic tool for *model selection* (that is, the problem of deciding which predictors to retain in a model) is the *F*-test for groups of predictors. We'll refer to (26.8) as the *full model* (in a more compact form)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_q X_q + \epsilon \text{ Full Model.} \quad (26.9)$$

For some  $p < q$  we have the *reduced model*

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon \text{ Reduced Model,} \quad (26.10)$$

that is, the reduced model is obtained from the full model by removing the final  $q - p$  predictors  $X_{p+1}, \dots, X_q$ . We say such models are *nested models*. The motivation here is to determine whether or not the predictive ability of the reduced model can be improved by adding these final predictors (there may be more than one of these). Of course, we have assumed that the predictors have been indexed appropriately.

Concluding that the full model is more predictive than the reduced model is equivalent to rejecting the hypothesis

$$H_o : \beta_{p+1} = \beta_{p+2} = \dots = \beta_q = 0$$

in favor of

$$H_a : \text{at least one of } \beta_{p+1}, \beta_{p+2}, \dots, \beta_q \text{ is not zero.}$$

The relevant *F* statistic is

$$F = \frac{(SSE_p - SSE_q)/(q - p)}{SSE_q/(n - (q + 1))}$$

Where  $SSE_q$  and  $SSE_p$  are the error sums of squares of the full and reduced model respectively. Under  $H_o$   $F$  has an *F*-distribution with  $q - p$  numerator degrees of freedom and  $n - (q + 1)$  denominator degrees of freedom, and so can be rejected at significance level  $\alpha$  if

$$F_{obs} \geq F_{q-p, n-(q+1), \alpha}.$$

It is important to note that we may set  $p = 0$ , in which case the reduce model is simply

$$Y = \beta_0 + \epsilon,$$

that is, responses are a random sample from  $N(\beta_0, \sigma^2)$  and are not related to any of the predictors (this is why  $\beta_0$  is usually retained in the model).

In fact, the  $F$ -statistic  $F = MSR/MSE$  given in the ANOVA table is the relevant test statistic, that is, it tests against the null hypothesis

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_q = 0,$$

as we have already seen.

The coefficient of multiple determination  $R^2$  must always be interpreted carefully, since it may be shown that its value is always larger for a full model than for a (nested) reduced model, even when the relevant coefficients are truly zero. This gives the often false impression that appending a new predictor to a model improves its predictive ability. Whether or not an increase in  $R^2$  is truly significant can be resolved by the appropriate  $F$ -test.

For this reason, we often use instead the *adjusted*  $R^2$ :

$$R_{adj}^2 = 1 - \frac{SSE/(n - (q + 1))}{SSTO/(n - 1)}.$$

This value, in a sense, is adjusted for the number of parameters, and permits a more accurate comparison between models with differing numbers of parameters, which need not be nested.

### 26.2.3 Example

Consider the following output for two regression models involving independent variables

*birthwt* (weight of infant at birth)

*headcirc* (head circumference of infant at birth)

*length* (length of infant at birth)

*toxemia* (= 1 if toxins present in blood, = 0 otherwise)

The objective is to estimate an infants prenatal or neonatal weight based on various measurements, which would be observable with a sonogram. The full model would be

$$birthwt = \beta_0 + \beta_1 \times headcirc + \beta_2 \times length + \beta_3 \times toxemia + \epsilon,$$

which has  $SSE(full) = 1647237.79$  with  $q = 3$  predictors. There is also a reduced model

$$birthwt = \beta_0 + \beta_1 \times headcirc + \epsilon,$$

with  $SSE(reduced) = 2611443.88$  with  $p = 1$  predictors. The sample size was  $n = 100$ . Model summaries are given below.

To test hypothesis

$$H_o : \beta_2 = \beta_3 = 0$$

against

$$H_a : \text{at least one of } \beta_2, \beta_3 \text{ is not zero}$$

we use  $F$ -statistic

$$\begin{aligned} F &= \frac{(SSE(\text{reduced}) - SSE(\text{full})) / 2}{SSE(\text{full}) / (100 - 4)} \\ &= \frac{(2611443.88 - 1647237.79) / 2}{1647237.79 / (100 - 4)} \\ &= 28.097. \end{aligned}$$

Under the null distribution  $H_o$ ,  $F$  has an  $F$  distribution with numerator and denominator degrees of freedom  $\nu_{\text{num}} = q-p = 2$  and  $\nu_{\text{den}} = n-(q+1) = 96$ . The  $p$ -value is very small, say  $P < 0.001$ , since we have critical value  $F_{2,96,0.001} = 7.43$ . So, the full model is more predictive than the reduced model.

Summary for reduced model:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
headcirc	1	4605298.87	4605298.87	172.82	0.0000
Residuals	98	2611443.88	26647.39		

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-1154.1087	172.1523	-6.70	0.0000
headcirc	85.1780	6.4793	13.15	0.0000

Summary for full model:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
headcirc	1	4605298.87	4605298.87	268.39	0.0000
length	1	889039.76	889039.76	51.81	0.0000
toxemia	1	75166.33	75166.33	4.38	0.0390
Residuals	96	1647237.79	17158.73		

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-1567.6048	148.7759	-10.54	0.0000
headcirc	48.3632	7.4349	6.50	0.0000
length	38.0639	5.2565	7.24	0.0000
toxemia	-67.9216	32.4518	-2.09	0.0390

## Chapter 27

# Linear Regression III

The chapter has two objectives. The first is to introduce R as a tool for statistical modeling. While this is carried out using linear regression, many of the methods are equally applicable to most other types of statistical models that one would encounter in an intermediate course on statistical methodology. For this reason, this chapter also introduces, mainly by example, some new modeling techniques which are of interest on their own.

### 27.1 Statistical Models

In a *statistical model* a random response  $Y$  is dependent on *predictors*  $X_1, X_2, \dots, X_m$ , in the sense that the distribution of  $Y$  depends on  $X_1, \dots, X_k$ . In many frequently used models, the relationship is given by

$$Y = \mu(X_1, X_2, \dots, X_m) + \epsilon, \quad \epsilon \sim N(0, \sigma^2), \quad (27.1)$$

or equivalently,

$$Y \sim N\left(\mu(X_1, X_2, \dots, X_m), \sigma^2\right).$$

ANOVA is a simple example of this. There is a single predictor  $X$ , which is a *factor*, or categorical variable, which assumes levels  $1, \dots, k$  (that is, there are  $k$  treatments, or groups). In this case, there are  $k$  means  $\mu_1, \dots, \mu_k$ , so that

$$\begin{aligned} Y &= \mu(X) + \epsilon \\ &= \mu_X + \epsilon \\ &= \mu_i + \epsilon \text{ if } X = i. \end{aligned}$$

Linear regression is a somewhat more complex example, but also conforms to Equation (27.1):

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_m X_m + \epsilon. \quad (27.2)$$

## 27.2 ANOVA as a Model in R

R supports a specialized notation for statistical models, based on the **formula** class. We have already seen a number of examples. In the following script, a data set consisting of a numerical vector **color.value** and a character vector **color.type** is created. There are 26 records, with equal numbers of “Red” and “Green” color types. The intention is that “Green” types tend to have higher values. The resulting plot is shown in Figure 27.1.

```
> par(mfrow=c(1,2),cex=1.0)
> color.value = rnorm(26,mean=rep(c(10,12),13))
> color.type = rep(c("Red","Green"),13)
> boxplot(color.value,ylab='Color Value')
> boxplot(color.value ~ color.type,ylab='Color Value')
```

The command `boxplot(color.value,ylab='Color Value')` creates a single boxplot of all the data, while the command `boxplot(color.value ~ color.type,ylab='Color Value')` creates side-by-side boxplots for each color type.

The expression `color.value ~ color.type` within the final boxplot command is an example of a **formula**, which takes the general form

$$\text{response} \sim \text{predictor expression}$$

It’s exact effect depends on the context. In it’s simplest form, as in the boxplot example of Figure 27.1, it separates a set of measurements by a group variable. However, it can also describe an analytical relationship between the response and multiple predictors.

The following script demonstrates the creation of a **formula** object, with the symbol `~` separating the response from the predictors:

```
> f1 = y ~ x
> f1
y ~ x
> class(f1)
[1] "formula"
```

The multiple regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \quad (27.3)$$

would be represented

```
> f1 = y ~ x1 + x2
> f1
y ~ x1 + x2
```

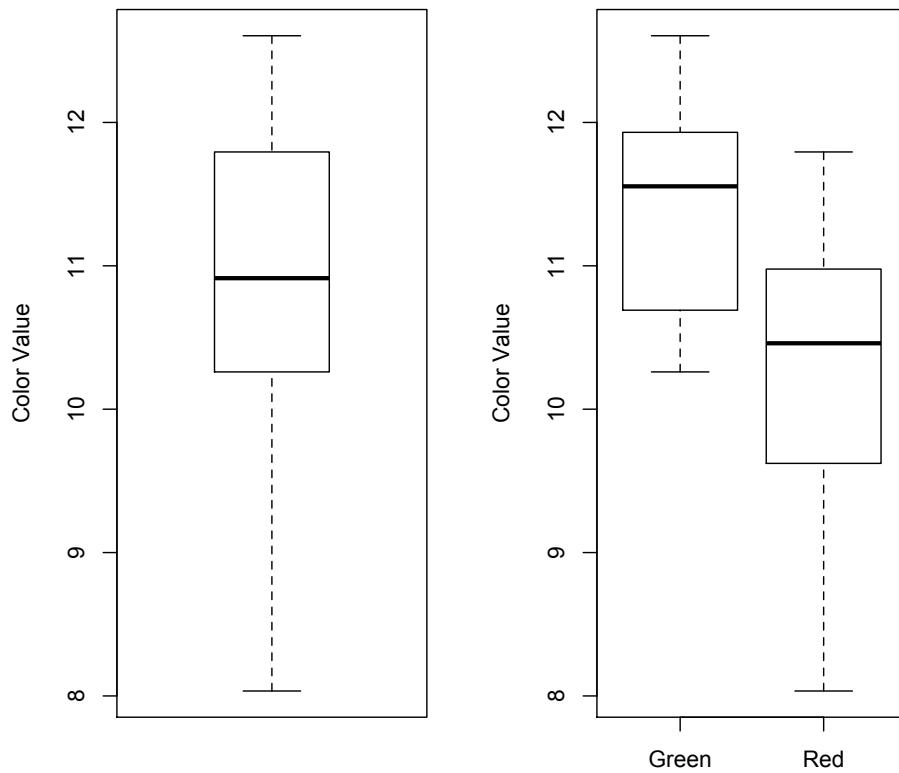


Figure 27.1: Example of the use of the `formula` class in the `boxplot`

We will make use later in the chapter of the *interaction term*, which is a predictor formed by taking the product of two or more other predictor terms. When interactions are present in a model, the predictors forming the interaction are referred to as *main effects*. In Equation (27.3) there is one possible interaction term  $X_1X_2$ , leading to regression equation

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2.$$

To specify all possible interactions and main effects involving two predictors the `*` operator can be used:

```
> f2 = y ~ x1 * x2
> f2
y ~ x1 * x2
> terms(f2)
y ~ x1 * x2
attr(,"variables")
list(y, x1, x2)
attr(,"factors")
  x1 x2 x1:x2
y   0   0     0
x1  1   0     1
x2  0   1     1
attr(,"term.labels")
[1] "x1"    "x2"    "x1:x2"
attr(,"order")
[1] 1 1 2
attr(,"intercept")
[1] 1
attr(,"response")
[1] 1
attr(,".Environment")
<environment: R_GlobalEnv>
```

Note that the function `terms()` is used to extract details of a formula.

If we wanted to include only the main effect for  $X_1$  and the interaction term, for example:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1 X_2$$

we could use the operator `:` to generate only the specified interactions.

```
> f3 = y ~ x1 + x1:x2
> f3
y ~ x1 + x1:x2
> terms(f3)
```

```

y ~ x1 + x1:x2
attr(,"variables")
list(y, x1, x2)
attr(,"factors")
  x1 x1:x2
y   0      0
x1  1      2
x2  0      1
attr(,"term.labels")
[1] "x1"     "x1:x2"
attr(,"order")
[1] 1 2
attr(,"intercept")
[1] 1
attr(,"response")
[1] 1
attr(,".Environment")
<environment: R_GlobalEnv>

```

The intercept term is implicitly included in a formula (but can be removed if needed). If we want to model to include only the intercept (for example, for model comparisons), we use the symbol 1:

```

> f4 = y ~ 1
> f4
y ~ 1
> terms(f4)
y ~ 1
attr(,"variables")
list(y)
attr(,"factors")
integer(0)
attr(,"term.labels")
character(0)
attr(,"order")
integer(0)
attr(,"intercept")
[1] 1
attr(,"response")
[1] 1
attr(,".Environment")
<environment: R_GlobalEnv>

```

This is a fairly in-depth topic, so we will discuss just the basics at first

(see `help(formula)` for more detail).

### 27.3 Linear Regression in R

First, we note that R comes with a set of example datasets, in a package called 'MASS'

```
> library(MASS)
> help(package=MASS)
...
Functions and datasets to support Venables and Ripley,
'Modern Applied Statistics with S' (4th edition, 2002).
...
```

One of these datasets is called `nlschools`:

```
> help(nlschools)
Description
```

Snijders and Bosker (1999) use as a running example a study of 2287 eighth-grade pupils (aged about 11) in 132 classes in 131 schools in the Netherlands. Only the variables used in our examples are supplied.

Usage

```
nlschools
Format
```

This data frame contains 2287 rows and the following columns:

```
lang
language test score.
```

```
IQ
verbal IQ.
```

```
class
class ID.
```

```
GS
class size: number of eighth-grade pupils recorded in the class
```

(there may be others: see COMB, and some may have been omitted with missing values).

SES

social-economic status of pupil's family.

COMB

were the pupils taught in a multi-grade class (0/1)? Classes which contained pupils from grades 7 and 8 are coded 1, but only eighth-graders were tested.

Source

Snijders, T. A. B. and Bosker, R. J. (1999) Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modelling. London: Sage.

References

Venables, W. N. and Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth edition. Springer.

The object `nlschools` is a data frame. You can verify that it is a list (a data frame is a list) using the `is.list()` function. The names of the variables is obtained using the `names()` command.

```
> is.list(nlschools)
[1] TRUE
> names(nlschools)
[1] "lang"    "IQ"      "class"   "GS"      "SES"     "COMB"
> nlschools[1:5,]
  lang   IQ class GS SES COMB
1  46 15.0  180 29  23    0
2  45 14.5  180 29  10    0
3  33  9.5  180 29  15    0
4  46 11.0  180 29  23    0
5  20  8.0  180 29  10    0
> dim(nlschools)
[1] 2287     6
>
```

There are 2287 rows and 6 columns.

Note that in the dataset `nlschols`, COMB is an *indicator variable*, that is, a variable that assumes only values 0,1 (or, a factor with two levels).

We can do a  $t$ -test to see if there is a significant difference in language test scores between students in multigrade classes, and those not in multigrade classes. We can use model notation, but we need to specify the data frame with the `data` option.

```
> t.test(lang ~ COMB, data=nlschools)

Welch Two Sample t-test

data: lang by COMB
t = 5.3849, df = 991.978, p-value = 9.052e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
1.540178 3.306355
sample estimates:
mean in group 0 mean in group 1
41.60133      39.17806
```

So, for  $\text{COMB} = 0$  (not in multigrade class) the estimated mean score is  $\hat{\mu}_0 = 41.6 \approx \mu_0$  and for  $\text{COMB} = 1$  (in multigrade class) the estimated mean score is  $\hat{\mu}_1 = 39.2 \approx \mu_1$ . There is a significant detrimental effect (about 2.4 points) on language test scores attributable to presence in multigrade class.

Next, suppose we consider regression model

$$\text{lang} = \beta_0 + \beta_1 \times \text{COMB} + \epsilon$$

Since `COMB` is an indicator variable, we can match the regression coefficients directly to the two group means:

$$\begin{aligned}\mu_0 &= \beta_0 \\ \mu_1 &= \beta_0 + \beta_1, \text{ with estimates} \\ \hat{\mu}_0 &= \hat{\beta}_0 \\ \hat{\mu}_1 &= \hat{\beta}_0 + \hat{\beta}_1.\end{aligned}$$

In R, regression fits can be calculated using the `lm()` function, using the model notation:

```
> fit = lm(lang ~ COMB, data=nlschools)
> summary(fit)

Call:
lm(formula = lang ~ COMB, data = nlschools)
```

Residuals:

	Min	1Q	Median	3Q	Max
-30.1781	-6.1781	0.8219	7.3987	18.8219	

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	41.6013	0.2196	189.472	< 2e-16 ***
COMB1	-2.4233	0.4187	-5.788	8.1e-09 ***
---				
Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’
	0.1 ‘ ’	1		

Residual standard error: 8.94 on 2285 degrees of freedom

Multiple R-squared: 0.01445, Adjusted R-squared: 0.01402

F-statistic: 33.5 on 1 and 2285 DF, p-value: 8.1e-09

We have coefficient estimates  $\hat{\beta}_0 = 41.6013$  and  $\hat{\beta}_1 = -2.4233$ , giving

$$\begin{aligned}\hat{\mu}_0 &= 41.6013 \\ \hat{\mu}_1 &= 41.6013 - 2.4233 = 39.178,\end{aligned}$$

which conform to the estimates we obtained above.

## 27.4 ANOVA and Linear Regression

Recall the ANOVA Example 16.1:

Coating	Sample	Sample Mean	Sum of Squares
A	10, 15, 8, 12, 15	$\bar{y}_1 = 12$	$\sum_{i=1}^5 (y_{1i} - \bar{y}_1)^2 = 38$
B	14, 18, 21, 15	$\bar{y}_2 = 17$	$\sum_{i=1}^4 (y_{2i} - \bar{y}_2)^2 = 30$
C	17, 16, 14, 15, 17, 15, 18	$\bar{y}_3 = 16$	$\sum_{i=1}^7 (y_{3i} - \bar{y}_3)^2 = 12$
D	12, 15, 17, 15, 16, 15	$\bar{y}_4 = 15$	$\sum_{i=1}^6 (y_{4i} - \bar{y}_4)^2 = 14$

We can create a data frame for the data in the following way:

```
> y1 = c(10, 15, 8, 12, 15)
> y2 = c(14, 18, 21, 15)
> y3 = c(17, 16, 14, 15, 17, 15, 18)
> y4 = c(12, 15, 17, 15, 16, 15)
> y = c(y1,y2,y3,y4)
> gr = c(rep("A",5), rep("B",4), rep("C",7), rep("D",6) )
>
> tapes.data = data.frame(y,gr)
```

```
> tapes.data
      y gr
1 10 A
2 15 A
3 8 A
4 12 A
5 15 A
6 14 B
7 18 B
8 21 B
9 15 B
10 17 C
11 16 C
12 14 C
13 15 C
14 17 C
15 15 C
16 18 C
17 12 D
18 15 D
19 17 D
20 15 D
21 16 D
22 15 D
>
```

The variable  $y$  contains the responses, with treatment groups indicators by the factor variable  $gr$ . As in the previous example, we can express the ANOVA model as a linear regression model using indicator variables:

$$Y = \beta_0 + \beta_1 \times I_B + \beta_2 \times I_C + \beta_3 \times I_D + \epsilon$$

where, for example,  $I_B$  is the indicator variable for treatment  $B$ . Note that we don't need (or want) an indicator variable for treatment  $A$ . The

coefficients can be related to the treatment means in the following way:

$$\begin{aligned}
 \mu_A &= \beta_0 \\
 \mu_B &= \beta_0 + \beta_1 \\
 \mu_C &= \beta_0 + \beta_2 \\
 \mu_D &= \beta_0 + \beta_3, \text{ with estimates} \\
 \hat{\mu}_A &= \hat{\beta}_0 \\
 \hat{\mu}_B &= \hat{\beta}_0 + \hat{\beta}_1 \\
 \hat{\mu}_C &= \hat{\beta}_0 + \hat{\beta}_2 \\
 \hat{\mu}_D &= \hat{\beta}_0 + \hat{\beta}_3.
 \end{aligned}$$

As can be seen, we only need indicators variables for 3 of the 4 treatments. To implement this using `lm()`, we could construct the indicator variables, but the same effect can be achieved using a factor variable.

```

> fit = lm(y ~ gr, data=tapes.data)
> summary(fit)

Call:
lm(formula = y ~ gr, data = tapes.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.00   -1.75    0.00    1.00    4.00 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 12.000     1.022  11.742 7.16e-10 ***
grB          5.000     1.533   3.262  0.00433 **  
grC          4.000     1.338   2.989  0.00787 **  
grD          3.000     1.384   2.168  0.04381 *   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.285 on 18 degrees of freedom
Multiple R-squared:  0.4198, Adjusted R-squared:  0.323 
F-statistic:  4.34 on 3 and 18 DF,  p-value: 0.01814

>

```

The coefficients match the model with

$$\begin{aligned}\hat{\beta}_0 &= 12.0 \\ \hat{\beta}_1 &= 5.0 \\ \hat{\beta}_2 &= 4.0 \\ \hat{\beta}_3 &= 3.0,\end{aligned}$$

and we can recreate the treatment mean estimates

$$\begin{aligned}\hat{\mu}_A &= \hat{\beta}_0 = 12.0 \\ \hat{\mu}_B &= \hat{\beta}_0 + \hat{\beta}_1 = 12.0 + 5.0 = 17.0 \\ \hat{\mu}_C &= \hat{\beta}_0 + \hat{\beta}_2 = 12.0 + 4.0 = 16.0 \\ \hat{\mu}_D &= \hat{\beta}_0 + \hat{\beta}_3 = 12.0 + 3.0 = 15.0.\end{aligned}$$

In addition, the  $F$  statistic  $F = 4.34$  is equivalent to that obtained by the ANOVA procedure, as is the  $F$  test for difference in means itself (compare to Example 16.1).

## 27.5 Residuals and `lm()`

The output of `lm()` is a list:

```
> names(fit)
[1] "coefficients"   "residuals"      "effects"       "rank"
"fitted.values"    "assign"        "qr"
"df.residual"      "contrasts"     "xlevels"
[11] "call"          "terms"         "model"
```

One of the components of this list is `residuals`, which is a vector of the residuals  $e_i = Y_i - \hat{Y}_i$ . We can, for example, examine the normality of the residuals with a normal quantile plot (Figure 22.1):

```
> qqnorm(fit$residual)
> qqline(fit$residual)
```

The quantile plot suggests that the assumption of normality is reasonable.

## 27.6 Interaction Terms

We'll return to the `nlschools` data, and fit the model

$$lang = \beta_0 + \beta_1 \times IQ + \epsilon$$

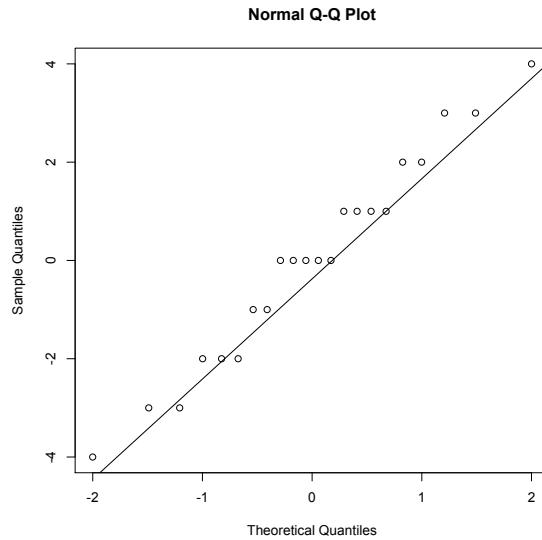


Figure 27.2: Normal quantile plot for Example 16.1 data.

The following script will fit the model, store the coefficients in a vector `cf`, do a scatter-plot of the independent against the dependent variable, then superimpose the actual regression line (Figure 22.2).

```
> fit = lm(lang ~ IQ, data = nlschools)
> summary(fit)

Call:
lm(formula = lang ~ IQ, data = nlschools)

Residuals:
    Min      1Q  Median      3Q     Max 
-28.7022 -4.3944  0.6056  5.2595 26.2212 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 9.52848   0.86682 10.99   <2e-16 ***
IQ          2.65390   0.07215 36.78   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.137 on 2285 degrees of freedom
Multiple R-squared:  0.3719, Adjusted R-squared:  0.3716 
F-statistic: 1353 on 1 and 2285 DF,  p-value: < 2.2e-16
```

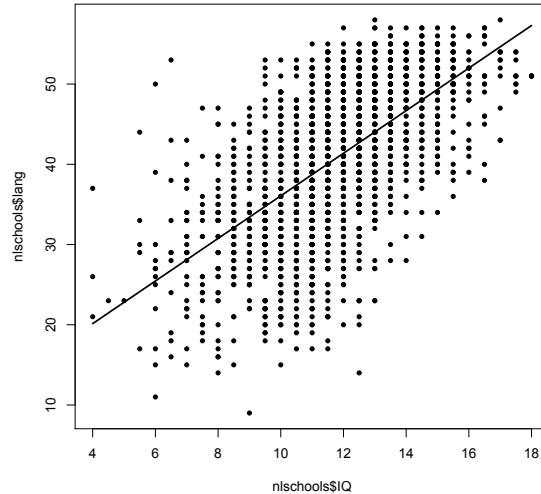


Figure 27.3: Regression fit for model  $lang = \beta_0 + \beta_1 \times IQ + \epsilon$

```

> cf = fit$coefficients
> cf
(Intercept)          IQ
  9.528484    2.653896
> range(nlschools$IQ)
[1]  4 18
> plot(nlschools$IQ, nlschools$lang, pch=20)
> lines(range(nlschools$IQ),
+ cf[1] + cf[2]*range(nlschools$IQ), lwd=2)
>

```

The following script can be used to produce diagnostic plots. Note that `par(mfrow=c(1,2))` permits two plots to appear on one window. The resulting residual plot and residual normal quantile plot appear in Figure 22.3.

```

>
> par(mfrow=c(1,2))
> plot(fit$fitted.values, fit$residuals, pch=20)
> lines(c(-100,100), c(0,0))
> qqnorm(fit$residuals)
> qqline(fit$residuals)
>

```

Note that these plots may also be obtained using `plot(fit)`. This type

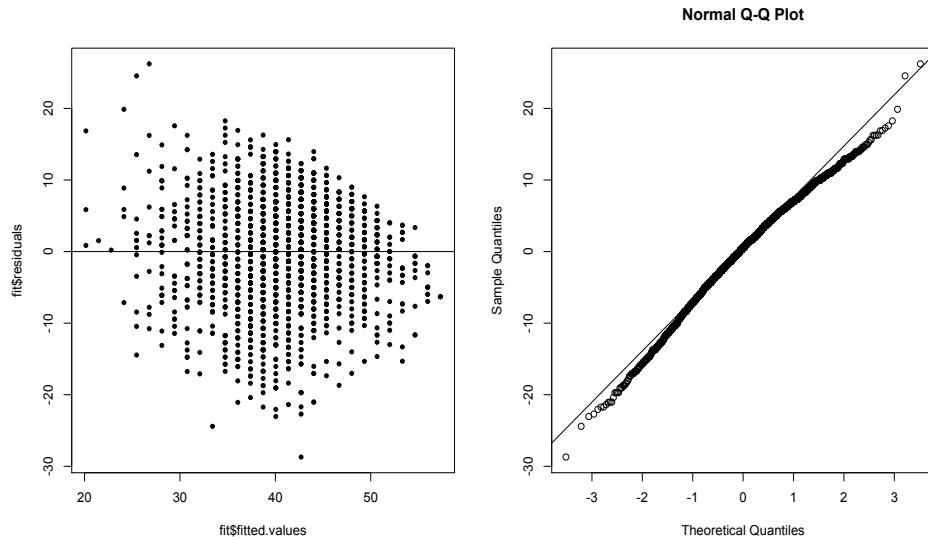


Figure 27.4: Residual plot and residual normal quantile plot for model  $lang = \beta_0 + \beta_1 \times IQ + \epsilon$

of feature is generally available for models in R.

Next, recall that the variable **COMB** had a significant effect on the language scores, so we may wish to introduce it into our model. First, remember to change the graphics window properties if needed (here, we only want one plot, so use `par(mfrow=c(1,1))`). We now have model

$$lang = \beta_0 + \beta_1 \times IQ + \beta_2 \times COMB + \epsilon$$

where **COMB** is an indicator variable. However, we can consider this as two linear regression models, one for multigrade classes, and one for single grade classes. We can then superimpose two fits on one plot, in particular  $y = \beta_0 + \beta_1 x$  for single grade classes, and  $y = (\beta_0 + \beta_2) + \beta_1 x$  for multigrade classes.

```
> par(mfrow=c(1,1))
> fit2 = lm(lang ~ IQ + COMB, data = nlschools)
> summary(fit2)

Call:
lm(formula = lang ~ IQ + COMB, data = nlschools)

Residuals:
    Min      1Q  Median      3Q     Max 
-29.55   -5.55   -1.55   13.55   29.55 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  10.5000    1.0000  10.500  <2e-16 ***
IQ           0.3000    0.0100  30.000  <2e-16 ***
COMB        -1.5000    0.5000  -3.000  0.0022 ** 
---
Signif. codes:  *** 0.001 ** 0.01 * 0.05 . 0.1  . 0.05  . 0.1  
```

```

-27.3890 -4.4989 0.5011 5.1841 25.6214

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.25824    0.87212   11.76 < 2e-16 ***
IQ           2.63390    0.07181   36.68 < 2e-16 ***
COMB1       -1.79296    0.33265   -5.39 7.77e-08 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.094 on 2284 degrees of freedom
Multiple R-squared: 0.3798, Adjusted R-squared: 0.3792
F-statistic: 699.2 on 2 and 2284 DF, p-value: < 2.2e-16

> cf = fit2$coefficients
> cf
(Intercept)          IQ          COMB1
  10.258240    2.633900   -1.792956
> plot(nlschools$IQ, nlschools$lang, pch=20)
> lines(range(nlschools$IQ),
  cf[1] + cf[2]*range(nlschools$IQ), lwd=2, col=2)
> lines(range(nlschools$IQ),
  cf[1] + cf[3] + cf[2]*range(nlschools$IQ), lwd=2, col=3)
> legend(14,20,
  legend=c("Multigrade class", "Single grade class"),
  lty=c(1,1), col=c(3,2))
>

```

Note here the use of the `col` option in `plot()` to color lines, thus distinguishing the groups. Also, the `lwd` option controls the width of the line, and `pch` defines the plotting symbol type. The `legend()` function is then used to add a legend. Compare the magnitude of the `COMB` effect of `-1.79296`, to the `COMB` effect obtained by the two sample mean comparison (also obtained by the simple regression fit  $lang = \beta_0 + \beta_1 \times COMB$ ) of `-2.4233`.

We may wonder, however, whether or not the slope of the regression line also differs by `COMB` group. To test for this, we can use *interaction terms*, which are simply products of other predictors. These are often very useful. When interactions are present, their components are referred to as *main effects*.

For example, we might fit model:

$$lang = \beta_0 + \beta_1 \times IQ + \beta_2 \times COMB + \beta_3 \times IQ \times COMB + \epsilon.$$

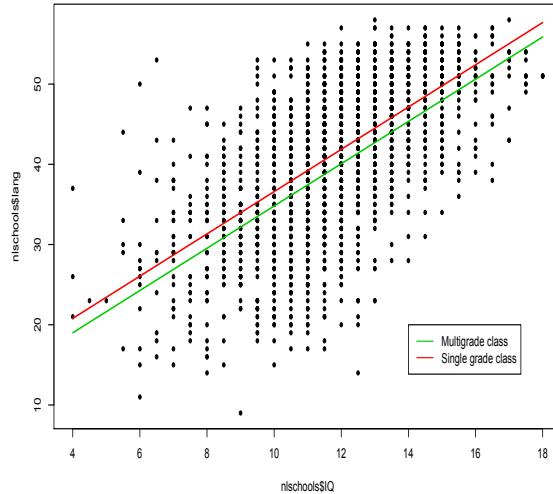


Figure 27.5: Regression fit for model  $lang = \beta_0 + \beta_1 \times IQ + \beta_2 \times COMB + \epsilon$

Here, both the intercept and slope can differ by `COMB` group:

$$lang = \beta_0 + \beta_1 \times IQ + \epsilon, \quad \text{for } COMB = 0$$

and

$$lang = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times IQ + \epsilon, \quad \text{for } COMB = 1.$$

In other words, the intercepts and slopes differ between the two `COMB` groups by  $\beta_2$  and  $\beta_3$  respectively, therefore such differences can be tested based on null hypotheses  $H_0 : \beta_2 = 0$  and  $H_0 : \beta_3 = 0$ .

We can fit this model using `lm()` (Figure 22.5). Interaction between two predictors are defined in model notion using the operator “`:`”. Alternatively, the operator “`*`” will introduce interactions and *main effects*.

```
> par(mfrow=c(1,1))
> fit2 = lm(lang ~ IQ + COMB + IQ:COMB, data = nlschools)
> summary(fit2)

Call:
lm(formula = lang ~ IQ + COMB + IQ:COMB, data = nlschools)

Residuals:
    Min      1Q  Median      3Q     Max 
-27.768 -4.484   0.473   5.153  24.646
```

```

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.4068    1.0228 12.131 < 2e-16 ***
IQ           2.4533    0.0847 28.966 < 2e-16 ***
COMB1       -9.2019   1.8875 -4.875 1.16e-06 ***
IQ:COMB1     0.6317    0.1584  3.987 6.90e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.071 on 2283 degrees of freedom
Multiple R-squared:  0.3841, Adjusted R-squared:  0.3832
F-statistic: 474.5 on 3 and 2283 DF,  p-value: < 2.2e-16

> cf = fit2$coefficients
> cf
(Intercept)          IQ          COMB1        IQ:COMB1
12.406772    2.453349   -9.201913    0.631680
> plot(nlschools$IQ, nlschools$lang, pch=20)
> lines(range(nlschools$IQ),
+ cf[1] + cf[2]*range(nlschools$IQ), lwd=2, col=2)
> lines(range(nlschools$IQ),
+ cf[1] + cf[3] + (cf[2]+cf[4])*range(nlschools$IQ), lwd=2, col=3)
> legend(14,20, legend=c("Multigrade class", "Single grade class"),
+ lty=c(1,1), col=c(3,2))

```

Here we see that language scores differ by COMB group, but also that this difference is larger among students with lower IQs. We say that IQ *interacts* with the COMB factor. This suggests that the performance of students with higher IQs may not be as influenced by external factors as other students are.

## 27.7 Polynomial Regression

The terms *linear* in *linear regression* nominally refers to the relationship between the predictors and the response. However, this has as much to do with the form of the inference as with any functional relationship. Suppose we have a model of the form

$$Y = 10 + 2.3x - 0.2x^2 + \epsilon \quad (27.4)$$

where  $\epsilon$  is the familiar error term. If we regard  $x$  as a single predictor, than Equation (27.4) conforms to (27.1) but not (27.2). On the other hand, we could also regard  $X_1 = x$  and  $X_2 = x^2$  as two distinct predictors, in which

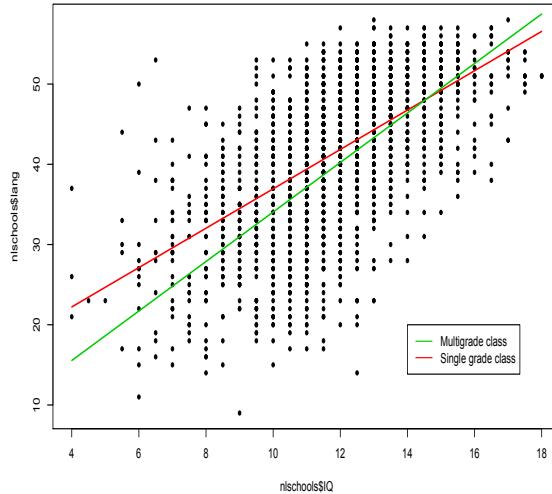


Figure 27.6: Regression fit for model  $lang = \beta_0 + \beta_1 \times IQ + \beta_2 \times COMB + \beta_3 \times IQ \times COMB + \epsilon$

case (27.4) conforms to both (27.1) and (27.2), with  $\beta_0 = 10$ ,  $\beta_1 = 2.3$  and  $\beta_2 = -0.2$ .

Fitting this type of model using multiple linear regression is referred to as *polynomial regression*. The methodology and inference remain exactly the same, as long as the linear structure of the inference is understood. In fact, introducing *quadratic terms* into a regression equation is a common method of both testing for nonlinear relationships between response and predictor, and for modeling such relationships when appropriate. We might first compare the full and reduced models

$$\begin{aligned} Y &= \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon, \\ Y &= \beta_0 + \beta_1 x + \epsilon. \end{aligned}$$

In this case, if a hypothesis test is able to reject the hypothesis  $H_0 : \beta_2 = 0$ , then the full model could be accepted and summarized. For more complex problems of this type the methods of Section 26.2.2 are available. We should note that somewhat more mathematically sophisticated methods exist for polynomial regression, which are implemented in R. One common practice is to use  $(x - \bar{x})^2$  instead of  $x^2$  as the quadratic term, particularly when a large range in  $x$  leads to a much larger range in  $x^2$ . Although the fitted values  $\hat{Y}$  will be identical using either form, the actual coefficient values will be different, and are usually more intuitively interpretable using the form  $(x - \bar{x})^2$ .

The quadratic term can be introduced into an `Rformula` object using the notation `I(x^2)`.

The following script simulates data from the model of Equation (27.4), using 19 equally spaced values for  $x$  ranging from 0 to 5.4. The error terms have standard deviation  $\sigma = 0.5$  (how can you tell this?). The model formula is created in object `lrform`, and used directly in function `lm()`. In general, elements of a formula can refer to columns in a data frame, which would then be explicitly reference using the `data` option in the `lm()` function, as shown in the examples using the `nlschools` data frame earlier in this chapter.

```
> f0 = function(x) {10 + 2.3*x - 0.2*x^2}
>
> x = seq(0,5.5,0.3)
> xsq = x^2
>
> plot(x,f0(x))
>
> mux = f0(x)
> y = mux + rnorm(length(x))/2
> plot(x,y,pch=20,cex=1)
> lrform = y~x + I(x^2)
> lrform
y ~ x + I(x^2)
> fit = lm(lrform)
> summary(fit)
```

Call:

```
lm(formula = lrform)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.56569	-0.25812	-0.05227	0.24923	0.75776

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	9.73262	0.24194	40.228	< 2e-16 ***
x	2.53668	0.20771	12.212	1.6e-09 ***
I(x^2)	-0.25088	0.03713	-6.758	4.6e-06 ***
---				
Signif. codes:	0 ***	0.001 **	0.01 *	0.05 . 0.1

Residual standard error: 0.3892 on 16 degrees of freedom  
Multiple R-squared: 0.9701, Adjusted R-squared: 0.9663

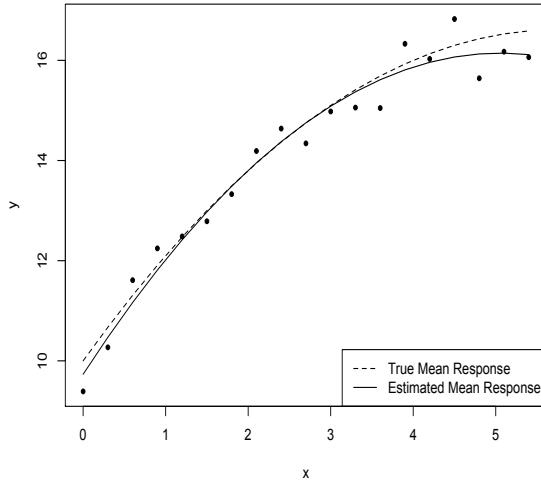


Figure 27.7: Data and linear model fit for polynomial regression example of Section 27.7.

```
F-statistic: 259.4 on 2 and 16 DF,  p-value: 6.417e-13
```

```
> lines(x,mux,lty=2)
> lines(x,predict(fit),lty=1)
> legend('bottomright',legend=c('True Mean Response','Estimated Mean Response'),col=c(2,1))
>
```

The resulting plot is shown in Figure 27.7. The fitted and true mean response curves are shown, along with the data points. The estimates  $\hat{\beta}_0 = 9.73262$ ,  $\hat{\beta}_1 = 2.53668$  and  $\hat{\beta}_2 = -0.25088$  are quite close to the true values  $\beta_0 = 10.0$ ,  $\beta_1 = 2.3$  and  $\beta_2 = -0.2$ .

## Chapter 28

# Classification and the Receiver Operator Characteristic (ROC) Curve

We already introduced in Chapter 5 a probabilistic model for the evaluation of a classifier, in the context of diagnostic testing. This was based on an application of Baye's theorem to the following events on a probability space:

$$\begin{aligned} O_+ &= \{ \text{positive outcome} \} \\ O_- &= \{ \text{negative outcome} \} \\ T_+ &= \{ \text{positive test outcome} \} \\ T_- &= \{ \text{negative test outcome} \}. \end{aligned}$$

We defined sensitivity and specificity as the following quantities:

$$\begin{aligned} sens &= P(T_+ | O_+) \\ spec &= P(T_- | O_-), \end{aligned}$$

and we may also define the *false positive rate* and *false negative rate* as

$$\begin{aligned} fpr &= P(T_+ | O_-) = 1 - spec \\ fnr &= P(T_- | O_+) = 1 - sens. \end{aligned}$$

These quantities are relevant in the evaluation phase of the development of a classifier. The ultimate goal is to maximize the positive predictive value (PPV) and negative predictive value (NPV), defined as

$$\begin{aligned} PPV &= P(O_+ | T_+) \\ NPV &= P(O_- | T_-), \end{aligned}$$

but to do so we need to test the classifier using subjects with known outcomes  $O_+$  and  $O_-$ , which gives *sens* and *spec*. We also need the prevalence of the outcome

$$\text{prev} = P(O_+),$$

with which Baye's theorem leads to

$$\begin{aligned} \text{PPV} &= P(O_+ | T_+) \\ &= \frac{P(T_+ | O_+)P(O_+)}{P(T_+ | O_+)P(O_+) + P(T_+ | O_-)P(O_-)} \\ &= \frac{\text{sens} \times \text{prev}}{\text{sens} \times \text{prev} + (1 - \text{spec}) \times (1 - \text{prev})} \end{aligned}$$

and

$$\begin{aligned} \text{NPV} &= P(O_- | T_-) \\ &= \frac{P(T_- | O_-)P(O_-)}{P(T_- | O_-)P(O_-) + P(T_- | O_+)P(O_+)} \\ &= \frac{\text{spec} \times (1 - \text{prev})}{\text{spec} \times (1 - \text{prev}) + (1 - \text{sens}) \times \text{prev}}. \end{aligned}$$

## 28.1 Classifiers Based on a Numerical Risk Score

The preceding section summarizes a probabilistic classification model for which the classifier can be reduced to two outcomes  $T_+$  and  $T_-$ . Of course, classifiers are often based on a numerical score. We can adopt the convention that higher scores can be interpreted as evidence in favor of a positive outcome  $O_+$  (if needed, reverse the score by multiplying by  $-1$ ). In this way, the numerical classifier can be interpreted as a *risk score*, with high risk implying greater probability (risk) of a positive outcome  $O_+$ .

To fix ideas, consider the `Melanoma` data set included in the `MASS` package:

```
> library(MASS)
> help(Melanoma)
```

Survival from Malignant Melanoma

### Description

The `Melanoma` data frame has data on 205 patients in Denmark with malignant melanoma.

### Usage

### Melanoma Format

This data frame contains the following columns:

**time**  
 survival time in days, possibly censored.

**status**  
 1 died from melanoma, 2 alive, 3 dead from other causes.

**sex**  
 1 = male, 0 = female.

**age**  
 age in years.

**year**  
 of operation.

**thickness**  
 tumour thickness in mm.

**ulcer**  
 1 = presence, 0 = absence.

### Source

P. K. Andersen, O. Borgan, R. D. Gill and  
 N. Keiding (1993) Statistical Models based on Counting  
 Processes. Springer.

We will investigate the possibility of using **thickness** (tumor thickness in mm) to predict death from melanoma. We have outcome **status**, which classifies the patient as dead from melanoma (**status** = 1); alive (**status** = 2); or dead from other causes (**tttstatus** = 3). We may remove from the analysis patients who died from other causes, leaving outcomes

$$\begin{aligned} O_+ &= \{\text{patient died from melanoma}\} \\ O_- &= \{\text{patient is still alive}\}. \end{aligned}$$

In practice, this type of analysis would take into account the observation

times of the patients, which may vary considerably. For example, a patient with outcome  $O_-$  may have only been observed for a short period of time, so that that negative outcome would be more difficult to interpret than an negative outcome which follows a longer observation period. With that caveat, we will accept survival as the outcome.

We have a quick first look at the data:

```
> names(Melanoma)
[1] "time"      "status"     "sex"       "age"       "year"
"thickness"   "ulcer"
> Melanoma[1:3,]
  time status sex age year thickness ulcer
1  10     3   1  76 1972      6.76      1
2  30     3   1  56 1968      0.65      0
3  35     2   1  41 1977      1.34      0
> is.factor(Melanoma$status)
[1] FALSE
>
```

Note that `status` is not a factor variable. So, to subset the data we use the command:

```
> Melanoma2 = Melanoma[Melanoma$status < 3,]
> dim(Melanoma2)
[1] 191   7
>
```

and use data frame `Melanoma2`. There are  $n = 191$  subjects remaining.

Next, look at boxplots of the variable `thickness` by outcome group (Figure 28.1):

```
> par(mfrow=c(1,1), mar=c(3,5,3,3), cex=1.1)
> boxplot(thickness ~ status, data = Melanoma2,
           names = c("Died", "Alive"), ylab="Tumor Thickness in mm.")
> for (i in 1:10) {lines(c(0,3), rep(i,2), col=4)}
```

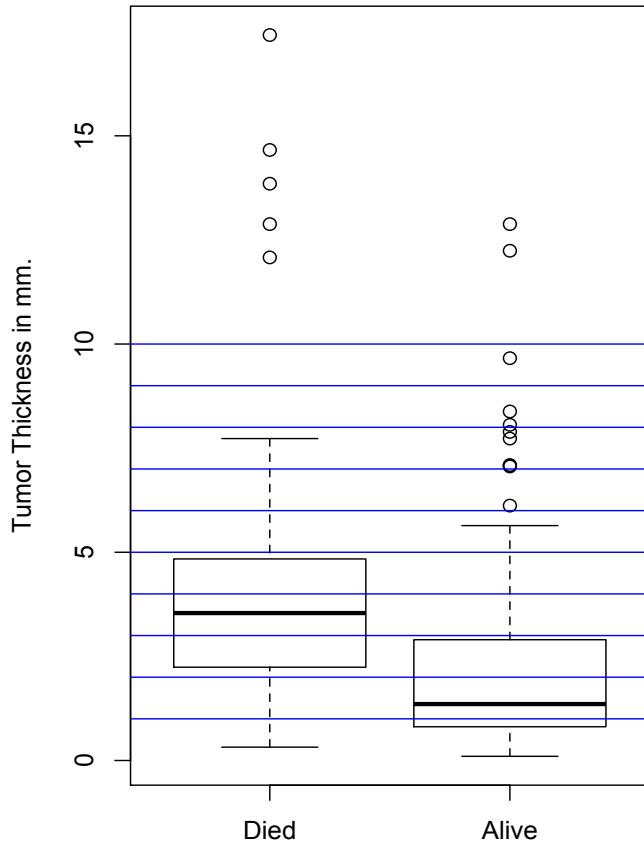


Figure 28.1: Boxplot of melanoma *thickness* variable (tumor thickness in mm) by survival outcome group.

We have superimposed lines (in blue) at *thickness* levels  $1, 2, \dots, 10$ . Clearly, death outcomes are associated with higher values of *thickness*, which can therefore be used as a risk score for melanoma mortality (higher values of *thickness* mean greater mortality risk). Suppose we select a *risk score threshold*  $T$ , possibly one of the blue lines. We may then define a positive test outcome as

$$T_+ = \{\text{thickness} \geq T\}. \quad (28.1)$$

This allows us to apply a classifier in an intuitive way, in the sense that if  $T_+$  occurs we predict  $O_+$ , and if  $T_- = T_+^c$  occurs we predict  $O_-$ .

Of course, this leaves open the problem of selecting  $T$ . If there were no overlap (that is, if the smallest risk score among the  $O_+$  group was larger

than the largest risk score among the  $O_-$  group) then the selection of  $T$  would be straightforward. If there was some  $T$  that is larger than all risk scores in the  $O_-$  group and smaller than all risk scores in the  $O_+$  group, we would use that threshold to define a positive test according to (28.1), which would yield  $sens = spec = 1$  that is, perfect classification (at least for this sample).

Of course, we don't usually expect this ideal. Suppose we consider the blue lines in Figure 28.1 as possible values for the threshold  $T$  used to define the positive test outcome  $T_+$  in (28.1). Clearly, for each value of  $T$  we will have false positives and false negatives, as long as within each group there are risk scores on both sides of the threshold.

It is instructive, however, to consider the limiting case. If  $T = 0$  (all risk scores are above 0), then the test outcome will be positive for *all* subjects in *both* groups. Since mortality is (correctly) predicted for all subjects in  $O_+$ , we have  $sens = 1$ . At the same time, mortality is (incorrectly) predicted for all subjects in  $O_-$ , so  $spec = 0$ . This is clearly not a satisfactory predictor. If  $T = 100$  (that is, a value larger than all observed risk scores) we (incorrectly) predict survival for all subjects in  $O_+$ , so that  $sens = 0$ . We also (correctly) predict survival for all subjects in  $O_-$ , so that  $spec = 1$ .

Clearly, we must find a balance between  $sens$  and  $spec$ . As  $T$  increases,  $sens$  decreases and  $spec$  increases. At this point, we can write an R function that calculates  $sens$  and  $spec$  for a given threshold, for a given data set. The function will have to input three things, namely, the threshold  $T$ , risk score  $score$  and the outcome groups  $gr$ . The variable  $gr$  will be a 0-1 numerical vector, with 1 corresponding to high risk. We assume  $score$  and  $gr$  are paired. Subjects with  $score \geq T$  are assigned positive test outcomes.

To estimate  $sens$  and  $spec$  from the data, we can use the following calculation:

$$\begin{aligned} sens &= \frac{P(O_+ \cap T_+)}{P(O_+)} = \frac{\text{Num subjects for which } score \geq T \text{ and } gr = 1}{\text{Num subjects for which } gr = 1} \\ spec &= \frac{P(O_- \cap T_-)}{P(O_-)} = \frac{\text{Num subjects for which } score < T \text{ and } gr = 0}{\text{Num subjects for which } gr = 0}. \end{aligned}$$

We therefore write the function:

```
> diag.thresh = function(thresh, score, gr) {
+   sens= sum( (score >= thresh) & (gr == 1) )/sum(gr == 1)
+   spec= sum( (score < thresh) & (gr == 0) )/sum(gr == 0)
+   ans = c(sens, spec)
+   names(ans) = c("Sensitivity", "Specificity")
+   return(ans)
+ }
```

&gt;

We can create our data variables for input

```
> gr = 1*(Melanoma2$status == 1)
> score = Melanoma2$thickness
> gr = gr[sort.list(score)]
> score = score[sort.list(score)]
```

Note that we have sorted the paired data using the `sort.list()` function. We can, for example, get the sensitivity associated with a threshold of  $T = 5$ :

```
> diag.thresh(5, score, gr)
Sensitivity Specificity
0.2456140 0.8955224
>
```

While the specificity is quite good ( $spec = 0.8955224$ ) the sensitivity would be, by most standards, too low ( $sens = 0.2456140$ ), and we would probably want to use a lower threshold for an actual application.

To see how the specificity and sensitivity vary with threshold  $T$ , we can create a loop to calculate a range of values for  $T$ , and create a simple table.

```
> diag.tab = NULL
> for (i in 1:10) {diag.tab =
  rbind(diag.tab,diag.thresh(i, score, gr))
+ }
> rownames(diag.tab) = paste("Threshold",1:10)
> colnames(diag.tab) = c("Sensitivity", "Specificity")
> diag.tab
      Sensitivity Specificity
Threshold 1    0.8947368  0.3507463
Threshold 2    0.7719298  0.6641791
Threshold 3    0.5964912  0.7611940
Threshold 4    0.3859649  0.8656716
Threshold 5    0.2456140  0.8955224
Threshold 6    0.1578947  0.9179104
Threshold 7    0.1403509  0.9253731
Threshold 8    0.0877193  0.9626866
Threshold 9    0.0877193  0.9776119
Threshold 10   0.0877193  0.9850746
>
```

A value of  $T$  in the range 2 to 3 would seem to offer a better balance of false positive and false negative rates.

## 28.2 ROC Curves

Of course, this type of analysis can be much more refined. First of all, we can input as threshold  $T$  all observed value of the risk score, obtaining much greater resolution than the previous table. To do this, we could use a `for` loop, but it's good to remember at this point that R permits vectorized operation. This would seem to suggest that if in the function call `diag.thresh(i, score, gr)` we substitute `score` for `i`, we would get the `sens, spec` values for all observed values of `score` in a single object. However, if we try this we get:

```
> diag.thresh(score, score, gr)
Sensitivity Specificity
      1          0
>
```

which is not what we wish. The problem lies in the fact that the other inputs are also vectors, leading to ambiguity. The problem may be fixed by using the `Vectorize()` function, which modifies an existing function by designating one or more of its inputs as the *vectorized* input as an option. The original function is evaluated for each element of the vectorized input. A new function is created in this way:

```
> help(Vectorize)
> diag.thresh.vect = Vectorize(diag.thresh, "thresh")
> temp = diag.thresh.vect(score,score,gr)
> dim(temp)
[1] 2 191
> sens = temp[,1]
> spec = temp[,2]
>
```

A new function `diag.thresh.vect()` has been created, which evaluates `diag.thresh()` separately for each element of the vector used as the first argument. The results are stored as a  $191 \times 2$  matrix, each column containing the values of `sens, spec` for each element of `score`.

At this point we are ready to plot an *ROC curve*, which is simply a plot of sensitivity (or true positive rate) against 1-specificity (or false positive rate) ('ROC' is an acronym for *receiver operating characteristic*). The script used to draw the plot is given below (Figure 28.2).

```
> auc = roc.area(class, gr)
> pv = wilcox.test(gr ~ class)$p.value
> par(mfrow=c(1,1), cex=1.1, oma = c(1,2,1,1))
> plot(1-spec, sens, xlab="false positive rate (1 - specificity)",
```

```

    ylab="true positive rate (sensitivity)", type = "s")
> title("ROC curve for prediction of melanoma
+         survival \n based on tumor thickness")
> lines(c(0,0),c(0,1),col=3)
> lines(c(0,1),c(1,1), col=3)
> lines(c(0,1), c(0,1))
> text(.7,.1, paste("AUC = ",signif(auc,3),",
+ P = ", signif(pv,3),sep=""))
>

```

First note the option `type = "s"` in the `plot()` function, which produces a step function type plot, which is appropriate for an ROC curve. Also, the control character "`\n`" may be used in the plot title to force a line break. In addition, an identity reference line has also been added to the plot, as well as green lines joining the points  $(0,0)$ ,  $(0,1)$  and  $(1,1)$ . The plot also gives two quantities,  $AUC$  as well as a  $p$ -value, which we now explain.

First, recall the “perfect” classifier discussed above, with sensitivity and specificity both equal to one. In this case, the ROC curve would coincide with the green lines of Figure 28.2. A highly accurate risk score would produce an ROC curve close to the green lines in some sense.

We next explain  $AUC$ . This is simply an acronym for *area under curve*. That is,  $AUC$  is defined as the area under the ROC curve. It may be shown that  $AUC$  is equal to the probability that a randomly chosen positive subject has a higher risk score than a randomly chosen negative subject. This can be given directly from the data:

$$AUC = \frac{\sum_{i \in \text{ve}} \sum_{j \in +\text{ve}} I\{score_j > score_i\} + 0.5 \times I\{score_j = score_i\}}{n_- \times n_+} \quad (28.2)$$

where  $n_-$ ,  $n_+$  are the number of negative and positive outcome subjects. Note that ties are assumed to be resolved randomly, hence the presence of the 0.5 factor in the numerator of (28.2). A function which calculates  $AUC$  is given below, and was used to calculate the value of  $AUC$  shown in figure Figure 28.2. This function is not, but could be, vectorized, as for the `diag.thresh.vect()` function given above.

```

> roc.area<-function(x,y) {
+
+   y0 = y[x==0]
+   y1 = y[x==1]
+
+   count<-0
+   for (i in 1:length(y0))
+     {count = count+sum(y1 > y0[i]) + 0.5*sum(y1 == y0[i])}

```

```
+ ans = count/(length(y0)*length(y1))
+ return(ans)
+
>
```

Suppose, in contrast to the perfect classifier indicated by the green line in Figure 28.2, that the risk score actually contains no information about the outcome. In this case, a randomly selected positive subject is equally likely to have a higher or lower risk score than a randomly selected negative subject. In this case we would expect  $AUC = 0.5$ , and the ROC curve would therefore lie on the identity. For this reason, the identity line is often included in an ROC curve graphic, and the degree to which the ROC curve lies above the identity gives a direct assessment of the predictive value of the risk score (the green lines are usually not given). What would you conclude if the ROC curve lay significantly *below* the identity?

Finally, we explain the  $p$ -value. It may be shown mathematically that the  $AUC$  is equivalent to the Wilcoxon rank sum statistic for a comparison of the risk score between the two outcome groups. This means the Wilcoxon rank sum test is interpretable as a test against the null hypothesis  $H_0 : AUC = 0.5$ . For this reason the  $p$ -value may be used to confirm that the risk score is significantly predictive of the outcome in a formal statistical sense.

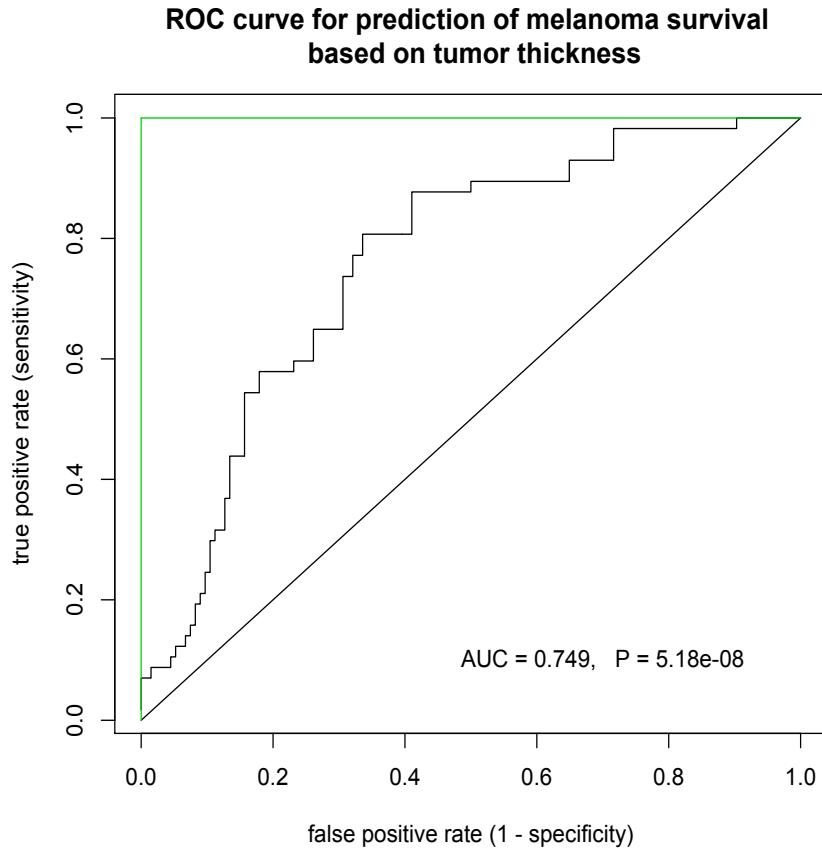


Figure 28.2: ROC curve for prediction of melanoma cancer survival based on melanoma *thickness* variable (tumor thickness in mm). The *AUC* is given, as well as the *p*-value for a Wilcoxon rank sum test for group homogeneity of risk score distributions. The green lines represent a “perfect” classifier, with sensitivity and specificity both equal to one. The diagonal identity line represents a noninformative risk score of *AUC* = 0.5.

# Chapter 29

## Simulation Methods

Much of the theory we have seen is based on approximations to the normal distribution. We either assume that the measurements we collect are normally distributed, or that the sample size is large enough for the Central Limit Theorem to apply to the test statistic. This applies also to test statistics with a  $\chi^2$  distribution used for categorical data, with the additional assumption that each category count is large enough.

There is very good reason to study normal-based theory, since much of it is provably optimal when the assumptions are satisfied, as they often are. However, these assumptions will often prove problematic, and even when they may hold, it might not be practical to verify them, for example, when a procedure is to be repeatedly applied to a large number of cases. This is generally the case in the analysis of, for example, gene expression data.

We then briefly consider two forms of simulation methods which do not rely on distributional assumptions, and which are generally applicable.

### 29.1 Permutation Test

A *permutation test* is a hypothesis test in which a null distribution is created by a random permutation of the data. Recall the data from Example 21.1,

```
X = 16.1 31.5 21.5 22.4 20.5 28.4 30.3 25.6 32.7 29.2 34.7  
Y = 4.41 6.81 5.26 5.99 5.92 6.14 6.84 5.87 7.03 6.89 7.87
```

for which the Pearson correlation coefficient was  $r_{obs} = 0.939$ . Suppose we randomly permute one of the variables (say,  $Y$ ), then recalculate  $r$ . We can generate a random permutation with the `sample()` function:

```
> sample(11)  
[1] 8 9 4 5 6 11 3 2 1 7 10  
>
```

We can then permute  $Y$ , then recalculate  $r$ :

```
> Yrandom = Y[sample(11)]
> X
[1] 16.1 31.5 21.5 22.4 20.5 28.4 30.3 25.6 32.7 29.2 34.7
> Yrandom
[1] 4.41 6.81 6.84 7.03 6.14 6.89 5.26 7.87 5.87 5.99 5.92
> cor(X,Y)
[1] 0.9388037
> cor(X,Yrandom)
[1] 0.1196821
>
```

The correlation of the permuted data,  $r^* = 0.1196821$ , is much smaller than the original  $r_{obs} = 0.939$ . This number is quite relevant, however. Under the null hypothesis  $H_0 : \rho = 0$ , there is no association between the paired variables  $X$  and  $Y$ . Therefore, if  $H_0$  is true, the observed sample correlation  $r_{obs}$  should be comparable to a correlation coefficient  $r^*$  produced by randomly permuting the data. This gives directly a test procedure that does not require any distribution assumptions. We can estimate the *null distribution* of  $r^*$  by repeatedly permuting the data, and then compared  $r_{obs}$  to this distribution, either by comparing it to a critical value of the null distribution, or by estimating the appropriate tail probability to obtain a  $p$ -value.

We first simulate  $r^*$   $N = 50,000$  times, and display the distribution in a histogram (Figure 29.1).

```
> r.perm = rep(NA,50000)
> for (i in 1:50000) {r.perm[i] = cor(X,Y[sample(11)])}
>
> hist(r.perm, nclass=25)
> lines(rep(0.735,2), c(0,5000), col=4)
> lines(rep(-0.735,2), c(0,5000), col=4)
> text(-0.735,5500,"r = -0.735")
> text(0.735,5500,"r = 0.735")
```

Recall from Example 21.1 that the critical value  $r_{\alpha/2}$  for a two-sided test against  $H_0 : \rho = 0$  ( $\alpha = 0.01$ ,  $n = 11$ ) was  $r_{\alpha/2} = 0.735$ , that is we reject  $H_0$  if  $|r_{obs}| \geq 0.735$ . This critical value is shown in Figure 29.1. This means that under the null distribution, the correlation coefficient satisfies:

$$P(|r| \geq 0.735 \mid \rho = 0) = 0.01.$$

We can estimate the same probability for  $r^*$  from the simulated null distribution:

```
> mean(abs(r.perm) >= 0.735)
[1] 0.00862
```

so that

$$P(|r^*| \geq 0.735 | \rho = 0) \approx 0.0082,$$

which is close to  $\alpha = 0.01$ . In fact, the level 95% margin of error of an estimate of a proportion  $p = 0.01$  with  $n = 50,000$  is

$$ME = 1.96 \sqrt{\frac{0.01 \times 0.99}{50000}} = 0.00087.$$

Judging from the margin of error, the tail probabilities for 0.735 is close to, but slightly less than,  $\alpha = 0.01$ . We can obtain a critical value for  $r_{obs}$  based on the distribution of  $r^*$  using the *quantile()* function:

```
> quantile(abs(r.perm), 0.99)
  99%
0.7257387
> quantile(r.perm, 0.995)
  99.5%
0.734525
> quantile(r.perm, 0.005)
  0.5%
-0.7181912
>
```

Note that we can obtain the  $\alpha/2 = 0.005$  critical value for  $r^*$  ( $r_{0.005}^* = 0.73425$ ), the lower tail  $1 - \alpha/2 = 0.995$  critical value ( $r_{1-0.005}^* = -0.7181912$ ), or the  $\alpha = 0.01$  critical value for  $|r^*|$  ( $|r^*|_{0.01} = 0.7257387$ ). If the null distribution is symmetric, which we would expect in this case, the critical value  $|r^*|_{0.01}$  should be used, since we would expect

$$|r^*|_\alpha \approx r_{\alpha/2}^* \approx -r_{1-\alpha/2}^*.$$

The  $p$ -value can be obtained by estimating the relevant tail probability of  $r_{obs}$ , in this case

$$P(r^* \geq r_{obs}), \quad P(r^* \leq r_{obs}), \quad \text{or } P(|r^*| \geq |r_{obs}|)$$

for an upper-tailed, lower-tailed or two sided test, respectively. However, it is usually the practice to add the observed statistics  $r_{obs}$  with the simulated values of  $r^*$  for this purpose, giving, for example:

$$P(|r^*| \geq |r_{obs}|) \approx \frac{\#\{|r^*| \geq |r_{obs}|\} + 1}{N + 1}. \quad (29.1)$$

This avoids  $p$ -values equal to zero, making the procedure somewhat conservative, although less so with increasing  $N$ . To assign a  $p$ -value to  $r_{obs} = 0.939$ , we can determine the numerator of (29.1) with the following command:

```
> sum(abs(r.perm) >= 0.939)
[1] 0
>
```

that is, no simulated value of  $r^*$  exceeds 0.939 in magnitude. This gives  $p$ -value

$$P \approx 1/50001 = 1.99996 \times 10^{-5}.$$

## 29.2 The Bootstrap Procedure

Suppose we are given a sample of size  $n = 10$ :

```
X = 36.1 16.1 16.7 32.7 33.9 21.8 15.5 26.0 37.8 18.6
```

A 95% confidence interval for the mean is given by

$$\bar{X} \pm t_{9,0.025} S / \sqrt{n} = 25.2 \pm 6.37 = (19.15, 31.89).$$

Remember that a confidence interval is a statement about a statistical method as well as a specific data set. If we could observe repeated samples collected under identical conditions, obtaining repeated observations of  $\bar{X}$ , we could observe the distribution of  $\bar{X}$  directly, and form inference statements accordingly, without the need to specify a distribution.

The *bootstrap procedure* is a method of simulating such samples, thus obtaining an estimated *sampling distribution* of, for example,  $\bar{X}$ , or any other statistic of interest. This is done by the simple device of sampling, *with replacement*, from the original sample (of size  $n$ ), a new sample of the same size  $n$ .

This can be done in R by the `sample()` function in the following way:

```
> n = 10
> sample(1:n, n, replace = TRUE)
[1] 1 6 4 2 3 5 5 8 8 3
>
```

(the command `sample.int(n, n, replace = TRUE)` will do the same thing). A *bootstrap sample* is then obtained by replacing the indices in the original sample:

```

> Xboot = X[sample(1:n, n, replace = TRUE)]
> X
[1] 36.1 16.1 16.7 32.7 33.9 21.8 15.5 26.0 37.8 18.6
> Xboot
[1] 33.9 21.8 16.1 37.8 36.1 15.5 16.7 32.7 21.8 16.1
> mean(X)
[1] 25.52
> mean(Xboot)
[1] 24.85
>

```

The bootstrap sample contains repeats, but we can still calculate most statistics for it. In this case, we get a new sample mean  $\bar{X}_{boot} = 24.85$  close to, but not exactly equal to, to original observed sample mean  $\bar{X}_{obs} = 25.52$ . As for the permutation procedure, we may then obtain a simulated sample, shown as a histogram in Figure 29.2.

```

> xbar.boot = rep(NA,50000)
> for (i in 1:50000)
  {xbar.boot[i] = mean(X[sample(1:n, n, replace = TRUE)])}
> hist(xbar.boot, nclass=25)
>

```

To obtain a 95% confidence interval, we need only obtain the 0.025 and 0.975 quantiles from the bootstrap sample,

```

> quantile(xbar.boot, c(0.025, 0.975))
2.5% 97.5%
20.40 30.81
>

```

yielding an estimated 95% confidence interval of (20.40, 30.81), which is quite close to the confidence interval (19.15, 31.89) obtained using the  $t$ -distribution above.

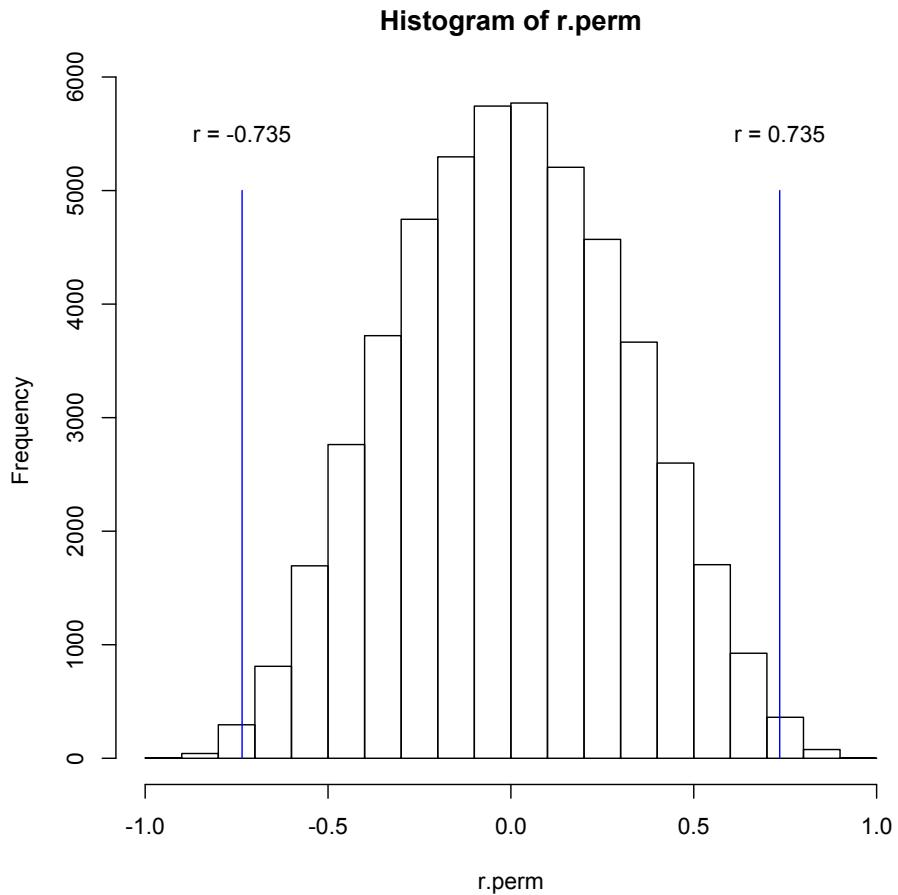


Figure 29.1: Histogram of 50,000 replications of  $r^*$ . The critical values  $r_{\alpha/2}$ ,  $-r_{\alpha/2}$  for a two-sided test against  $H_0 : \rho = 0$  ( $\alpha = 0.01$ ,  $n = 11$ ) are superimposed.

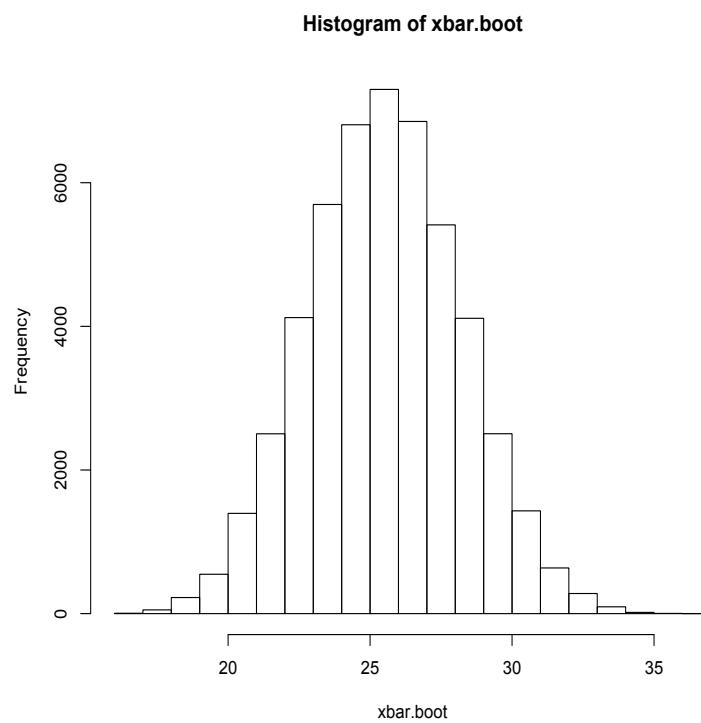


Figure 29.2: Histogram of 50,000 replications of  $\bar{X}_{boot}$ .

# **Appendix A**

# **Distribution Tables**

Table A.1: **Standard Normal Curve Areas I.** The table entry is the probability that a standard normal random variable is less than or equal to  $z$ .

$z$	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.001	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.001	0.001
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.002	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.003	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.004	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.006	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.008	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.011
-2.1	0.0179	0.0174	0.017	0.0166	0.0162	0.0158	0.0154	0.015	0.0146	0.0143
-2	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.025	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.063	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.102	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.123	0.121	0.119	0.117
-1	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.166	0.1635	0.1611
-0.8	0.2119	0.209	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.242	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.305	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.281	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.33	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.352	0.3483
-0.2	0.4207	0.4168	0.4129	0.409	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0	0.5	0.496	0.492	0.488	0.484	0.4801	0.4761	0.4721	0.4681	0.4641

**Table A.2: Standard Normal Curve Areas II.** The table entry is the probability that a standard normal random variable is less than or equal to

Table A.3: **Critical values for  $t$ -distribution.** Table entry is the  $\alpha$  critical value  $t_{df,\alpha}$  for a  $t$ -distribution with  $df$  degrees of freedom.

$df$	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005	0.00025
1	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619	1273.239
2	1.886	2.92	4.303	6.965	9.925	14.089	22.327	31.599	44.705
3	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924	16.326
4	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.61	10.306
5	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869	7.976
6	1.44	1.943	2.447	3.143	3.707	4.317	5.208	5.959	6.788
7	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408	6.082
8	1.397	1.86	2.306	2.896	3.355	3.833	4.501	5.041	5.617
9	1.383	1.833	2.262	2.821	3.25	3.69	4.297	4.781	5.291
10	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587	5.049
11	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437	4.863
12	1.356	1.782	2.179	2.681	3.055	3.428	3.93	4.318	4.716
13	1.35	1.771	2.16	2.65	3.012	3.372	3.852	4.221	4.597
14	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.14	4.499
15	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073	4.417
16	1.337	1.746	2.12	2.583	2.921	3.252	3.686	4.015	4.346
17	1.333	1.74	2.11	2.567	2.898	3.222	3.646	3.965	4.286
18	1.33	1.734	2.101	2.552	2.878	3.197	3.61	3.922	4.233
19	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883	4.187
20	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.85	4.146
21	1.323	1.721	2.08	2.518	2.831	3.135	3.527	3.819	4.11
22	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792	4.077
23	1.319	1.714	2.069	2.5	2.807	3.104	3.485	3.768	4.047
24	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745	4.021
25	1.316	1.708	2.06	2.485	2.787	3.078	3.45	3.725	3.996
26	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707	3.974
27	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.69	3.954
28	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674	3.935
29	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659	3.918
30	1.31	1.697	2.042	2.457	2.75	3.03	3.385	3.646	3.902
40	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551	3.788
50	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496	3.723
60	1.296	1.671	2	2.39	2.66	2.915	3.232	3.46	3.681
70	1.294	1.667	1.994	2.381	2.648	2.899	3.211	3.435	3.651
80	1.292	1.664	1.99	2.374	2.639	2.887	3.195	3.416	3.629
90	1.291	1.662	1.987	2.368	2.632	2.878	3.183	3.402	3.612
100	1.29	1.66	1.984	2.364	2.626	2.871	3.174	3.39	3.598
110	1.289	1.659	1.982	2.361	2.621	2.865	3.166	3.381	3.587
120	1.289	1.658	1.98	2.358	2.617	2.86	3.16	3.373	3.578
Inf	1.282	1.645	1.96	2.326	2.576	2.807	3.09	3.291	3.481

Table A.4: **Critical values for  $\chi^2$ -distribution.** Table entry is the  $\alpha$  critical value  $\chi_{df,\alpha}^2$  for a  $\chi^2$ -distribution with  $df$  degrees of freedom.

$df$	$\alpha$									
	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
1	0	0	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.01	0.02	0.051	0.103	0.211	4.605	5.991	7.378	9.21	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.86
5	0.412	0.554	0.831	1.145	1.61	9.236	11.07	12.833	15.086	16.75
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.69	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.18	2.733	3.49	13.362	15.507	17.535	20.09	21.955
9	1.735	2.088	2.7	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.94	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.92	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.3
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.66	5.629	6.571	7.79	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.39	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.26	9.591	10.851	12.443	28.412	31.41	34.17	37.566	39.997
21	8.034	8.897	10.283	11.591	13.24	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.26	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.98	45.559
25	10.52	11.524	13.12	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.16	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.29
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
31	14.458	15.655	17.539	19.281	21.434	41.422	44.985	48.232	52.191	55.003
32	15.134	16.362	18.291	20.072	22.271	42.585	46.194	49.48	53.486	56.328
33	15.815	17.074	19.047	20.867	23.11	43.745	47.4	50.725	54.776	57.648
34	16.501	17.789	19.806	21.664	23.952	44.903	48.602	51.966	56.061	58.964
35	17.192	18.509	20.569	22.465	24.797	46.059	49.802	53.203	57.342	60.275
36	17.887	19.233	21.336	23.269	25.643	47.212	50.998	54.437	58.619	61.581
37	18.586	19.96	22.106	24.075	26.492	48.363	52.192	55.668	59.893	62.883
38	19.289	20.691	22.878	24.884	27.343	49.513	53.384	56.896	61.162	64.181
39	19.996	21.426	23.654	25.695	28.196	50.66	54.572	58.12	62.428	65.476
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
41	21.421	22.906	25.215	27.326	29.907	52.949	56.942	60.561	64.95	68.053
42	22.138	23.65	25.999	28.144	30.765	54.09	58.124	61.777	66.206	69.336
43	22.859	24.398	26.785	28.965	31.625	55.23	59.304	62.99	67.459	70.616
44	23.584	25.148	27.575	29.787	32.487	56.369	60.481	64.201	68.71	71.893
45	24.311	25.901	28.366	30.612	33.35	57.505	61.656	65.41	69.957	73.166
46	25.041	26.657	29.16	31.439	34.215	58.641	62.83	66.617	71.201	74.437
47	25.775	27.416	29.956	32.268	35.081	59.774	64.001	67.821	72.443	75.704
48	26.511	28.177	30.755	33.098	35.949	60.907	65.171	69.023	73.683	76.969
49	27.249	28.941	31.555	33.93	36.818	62.038	66.339	70.222	74.919	78.231
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.42	76.154	79.49
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.54	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169
150	109.142	112.668	117.985	122.692	128.275	172.581	179.581	185.8	193.208	198.36
200	152.241	156.432	162.728	168.279	174.835	226.021	233.994	241.058	249.445	255.264
250	196.161	200.939	208.098	214.392	221.806	279.05	287.882	295.689	304.94	311.346
300	240.663	245.972	253.912	260.878	269.068	331.789	341.395	349.874	359.906	366.844
350	285.608	291.406	300.064	307.648	316.55	384.306	394.626	403.723	414.474	421.9
400	330.903	337.155	346.482	354.641	364.207	436.649	447.632	457.305	468.724	476.606
450	376.483	383.163	393.118	401.817	412.007	488.849	500.456	510.67	522.717	531.026
500	422.303	429.388	439.936	449.147	459.926	540.93	553.127	563.852	576.493	585.207

Table A.5: **Critical values for  $F$ -distribution I.** Table entry is the  $\alpha$  critical value  $\chi_{df_1, df_2, \alpha}^2$  for an  $F$ -distribution with  $df_1$  numerator degrees of freedom and  $df_2$  denominator degrees of freedom.

$df_2$	$\alpha$	1	2	3	4	5	6	7	8	9
		$df_1$								
1	0.1	39.86	49.5	53.59	55.83	57.24	58.2	58.91	59.44	59.86
1	0.05	161.45	199.5	215.71	224.58	230.16	233.99	236.77	238.88	240.54
1	0.025	647.79	799.5	864.16	899.58	921.85	937.11	948.22	956.66	963.28
1	0.01	4052.18	4999.5	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47
1	0.005	16210.72	19999.5	21614.74	22499.58	23055.8	23437.11	23714.57	23925.41	24091
2	0.1	8.53	9	9.16	9.24	9.29	9.33	9.35	9.37	9.38
2	0.05	18.51	19	19.16	19.25	19.3	19.33	19.35	19.37	19.38
2	0.025	38.51	39	39.17	39.25	39.3	39.33	39.36	39.37	39.39
2	0.01	98.5	99	99.17	99.25	99.3	99.33	99.36	99.37	99.39
2	0.005	198.5	199	199.17	199.25	199.3	199.33	199.36	199.37	199.39
3	0.1	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24
3	0.05	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
3	0.025	17.44	16.04	15.44	15.1	14.88	14.73	14.62	14.54	14.47
3	0.01	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
3	0.005	55.55	49.8	47.47	46.19	45.39	44.84	44.43	44.13	43.88
4	0.1	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94
4	0.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6
4	0.025	12.22	10.65	9.98	9.6	9.36	9.2	9.07	8.98	8.9
4	0.01	21.2	18	16.69	15.98	15.52	15.21	14.98	14.8	14.66
4	0.005	31.33	26.28	24.26	23.15	22.46	21.97	21.62	21.35	21.14
5	0.1	4.06	3.78	3.62	3.52	3.45	3.4	3.37	3.34	3.32
5	0.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
5	0.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68
5	0.01	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
5	0.005	22.78	18.31	16.53	15.56	14.94	14.51	14.2	13.96	13.77
6	0.1	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96
6	0.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.1
6	0.025	8.81	7.26	6.6	6.23	5.99	5.82	5.7	5.6	5.52
6	0.01	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.1	7.98
6	0.005	18.63	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.39
7	0.1	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72
7	0.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
7	0.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.9	4.82
7	0.01	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
7	0.005	16.24	12.4	10.88	10.05	9.52	9.16	8.89	8.68	8.51
8	0.1	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56
8	0.05	5.32	4.46	4.07	3.84	3.69	3.58	3.5	3.44	3.39
8	0.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36
8	0.01	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
8	0.005	14.69	11.04	9.6	8.81	8.3	7.95	7.69	7.5	7.34
9	0.1	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44
9	0.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
9	0.025	7.21	5.71	5.08	4.72	4.48	4.32	4.2	4.1	4.03
9	0.01	10.56	8.02	6.99	6.42	6.06	5.8	5.61	5.47	5.35
9	0.005	13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.54

Table A.6: **Critical values for  $F$ -distribution II.** Table entry is the  $\alpha$  critical value  $\chi_{df_1, df_2, \alpha}^2$  for an  $F$ -distribution with  $df_1$  numerator degrees of freedom and  $df_2$  denominator degrees of freedom.

$df_2$	$\alpha$	10	12	15	20	24	30	60	120	Inf
1	0.1	60.19	60.71	61.22	61.74	62	62.26	62.79	63.06	63.33
1	0.05	241.88	243.91	245.95	248.01	249.05	250.1	252.2	253.25	254.31
1	0.025	968.63	976.71	984.87	993.1	997.25	1001.41	1009.8	1014.02	1018.26
1	0.01	6055.85	6106.32	6157.28	6208.73	6234.63	6260.65	6313.03	6339.39	6365.86
1	0.005	24224.49	24426.37	24630.21	24835.97	24939.57	25043.63	25253.14	25358.57	25464.46
2	0.1	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.48	9.49
2	0.05	19.4	19.41	19.43	19.45	19.45	19.46	19.48	19.49	19.5
2	0.025	39.4	39.41	39.43	39.45	39.46	39.46	39.48	39.49	39.5
2	0.01	99.4	99.42	99.43	99.45	99.46	99.47	99.48	99.49	99.5
2	0.005	199.4	199.42	199.43	199.45	199.46	199.47	199.48	199.49	199.5
3	0.1	5.23	5.22	5.2	5.18	5.18	5.17	5.15	5.14	5.13
3	0.05	8.79	8.74	8.7	8.66	8.64	8.62	8.57	8.55	8.53
3	0.025	14.42	14.34	14.25	14.17	14.12	14.08	13.99	13.95	13.9
3	0.01	27.23	27.05	26.87	26.69	26.6	26.5	26.32	26.22	26.13
3	0.005	43.69	43.39	43.08	42.78	42.62	42.47	42.15	41.99	41.83
4	0.1	3.92	3.9	3.87	3.84	3.83	3.82	3.79	3.78	3.76
4	0.05	5.96	5.91	5.86	5.8	5.77	5.75	5.69	5.66	5.63
4	0.025	8.84	8.75	8.66	8.56	8.51	8.46	8.36	8.31	8.26
4	0.01	14.55	14.37	14.2	14.02	13.93	13.84	13.65	13.56	13.46
4	0.005	20.97	20.7	20.44	20.17	20.03	19.89	19.61	19.47	19.32
5	0.1	3.3	3.27	3.24	3.21	3.19	3.17	3.14	3.12	3.1
5	0.05	4.74	4.68	4.62	4.56	4.53	4.5	4.43	4.4	4.36
5	0.025	6.62	6.52	6.43	6.33	6.28	6.23	6.12	6.07	6.02
5	0.01	10.05	9.89	9.72	9.55	9.47	9.38	9.2	9.11	9.02
5	0.005	13.62	13.38	13.15	12.9	12.78	12.66	12.4	12.27	12.14
6	0.1	2.94	2.9	2.87	2.84	2.82	2.8	2.76	2.74	2.72
6	0.05	4.06	4	3.94	3.87	3.84	3.81	3.74	3.7	3.67
6	0.025	5.46	5.37	5.27	5.17	5.12	5.07	4.96	4.9	4.85
6	0.01	7.87	7.72	7.56	7.4	7.31	7.23	7.06	6.97	6.88
6	0.005	10.25	10.03	9.81	9.59	9.47	9.36	9.12	9	8.88
7	0.1	2.7	2.67	2.63	2.59	2.58	2.56	2.51	2.49	2.47
7	0.05	3.64	3.57	3.51	3.44	3.41	3.38	3.3	3.27	3.23
7	0.025	4.76	4.67	4.57	4.47	4.41	4.36	4.25	4.2	4.14
7	0.01	6.62	6.47	6.31	6.16	6.07	5.99	5.82	5.74	5.65
7	0.005	8.38	8.18	7.97	7.75	7.64	7.53	7.31	7.19	7.08
8	0.1	2.54	2.5	2.46	2.42	2.4	2.38	2.34	2.32	2.29
8	0.05	3.35	3.28	3.22	3.15	3.12	3.08	3.01	2.97	2.93
8	0.025	4.3	4.2	4.1	4	3.95	3.89	3.78	3.73	3.67
8	0.01	5.81	5.67	5.52	5.36	5.28	5.2	5.03	4.95	4.86
8	0.005	7.21	7.01	6.81	6.61	6.5	6.4	6.18	6.06	5.95
9	0.1	2.42	2.38	2.34	2.3	2.28	2.25	2.21	2.18	2.16
9	0.05	3.14	3.07	3.01	2.94	2.9	2.86	2.79	2.75	2.71
9	0.025	3.96	3.87	3.77	3.67	3.61	3.56	3.45	3.39	3.33
9	0.01	5.26	5.11	4.96	4.81	4.73	4.65	4.48	4.4	4.31
9	0.005	6.42	6.23	6.03	5.83	5.73	5.62	5.41	5.3	5.19

Table A.7: **Critical values for  $F$ -distribution III.** Table entry is the  $\alpha$  critical value  $\chi^2_{df_1, df_2, \alpha}$  for an  $F$ -distribution with  $df_1$  numerator degrees of freedom and  $df_2$  denominator degrees of freedom.

$df_2$	$\alpha$	1	2	3	4	5	6	7	8	9
10	0.1	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35
10	0.05	4.96	4.1	3.71	3.48	3.33	3.22	3.14	3.07	3.02
10	0.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78
10	0.01	10.04	7.56	6.55	5.99	5.64	5.39	5.2	5.06	4.94
10	0.005	12.83	9.43	8.08	7.34	6.87	6.54	6.3	6.12	5.97
12	0.1	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21
12	0.05	4.75	3.89	3.49	3.26	3.11	3	2.91	2.85	2.8
12	0.025	6.55	5.1	4.47	4.12	3.89	3.73	3.61	3.51	3.44
12	0.01	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.5	4.39
12	0.005	11.75	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.2
15	0.1	3.07	2.7	2.49	2.36	2.27	2.21	2.16	2.12	2.09
15	0.05	4.54	3.68	3.29	3.06	2.9	2.79	2.71	2.64	2.59
15	0.025	6.2	4.77	4.15	3.8	3.58	3.41	3.29	3.2	3.12
15	0.01	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4	3.89
15	0.005	10.8	7.7	6.48	5.8	5.37	5.07	4.85	4.67	4.54
20	0.1	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2	1.96
20	0.05	4.35	3.49	3.1	2.87	2.71	2.6	2.51	2.45	2.39
20	0.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84
20	0.01	8.1	5.85	4.94	4.43	4.1	3.87	3.7	3.56	3.46
20	0.005	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96
24	0.1	2.93	2.54	2.33	2.19	2.1	2.04	1.98	1.94	1.91
24	0.05	4.26	3.4	3.01	2.78	2.62	2.51	2.42	2.36	2.3
24	0.025	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.7
24	0.01	7.82	5.61	4.72	4.22	3.9	3.67	3.5	3.36	3.26
24	0.005	9.55	6.66	5.52	4.89	4.49	4.2	3.99	3.83	3.69
30	0.1	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85
30	0.05	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
30	0.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57
30	0.01	7.56	5.39	4.51	4.02	3.7	3.47	3.3	3.17	3.07
30	0.005	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45
60	0.1	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74
60	0.05	4	3.15	2.76	2.53	2.37	2.25	2.17	2.1	2.04
60	0.025	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33
60	0.01	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
60	0.005	8.49	5.79	4.73	4.14	3.76	3.49	3.29	3.13	3.01
120	0.1	2.75	2.35	2.13	1.99	1.9	1.82	1.77	1.72	1.68
120	0.05	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96
120	0.025	5.15	3.8	3.23	2.89	2.67	2.52	2.39	2.3	2.22
120	0.01	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56
120	0.005	8.18	5.54	4.5	3.92	3.55	3.28	3.09	2.93	2.81
Inf	0.1	2.71	2.3	2.08	1.94	1.85	1.77	1.72	1.67	1.63
Inf	0.05	3.84	3	2.6	2.37	2.21	2.1	2.01	1.94	1.88
Inf	0.025	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11
Inf	0.01	6.63	4.61	3.78	3.32	3.02	2.8	2.64	2.51	2.41
Inf	0.005	7.88	5.3	4.28	3.72	3.35	3.09	2.9	2.74	2.62

Table A.8: **Critical values for  $F$ -distribution IV.** Table entry is the  $\alpha$  critical value  $\chi^2_{df_1, df_2, \alpha}$  for an  $F$ -distribution with  $df_1$  numerator degrees of freedom and  $df_2$  denominator degrees of freedom.

$df_2$	$\alpha$	10	12	15	20	24	30	60	120	Inf
10	0.1	2.32	2.28	2.24	2.2	2.18	2.16	2.11	2.08	2.06
10	0.05	2.98	2.91	2.85	2.77	2.74	2.7	2.62	2.58	2.54
10	0.025	3.72	3.62	3.52	3.42	3.37	3.31	3.2	3.14	3.08
10	0.01	4.85	4.71	4.56	4.41	4.33	4.25	4.08	4	3.91
10	0.005	5.85	5.66	5.47	5.27	5.17	5.07	4.86	4.75	4.64
12	0.1	2.19	2.15	2.1	2.06	2.04	2.01	1.96	1.93	1.9
12	0.05	2.75	2.69	2.62	2.54	2.51	2.47	2.38	2.34	2.3
12	0.025	3.37	3.28	3.18	3.07	3.02	2.96	2.85	2.79	2.72
12	0.01	4.3	4.16	4.01	3.86	3.78	3.7	3.54	3.45	3.36
12	0.005	5.09	4.91	4.72	4.53	4.43	4.33	4.12	4.01	3.9
15	0.1	2.06	2.02	1.97	1.92	1.9	1.87	1.82	1.79	1.76
15	0.05	2.54	2.48	2.4	2.33	2.29	2.25	2.16	2.11	2.07
15	0.025	3.06	2.96	2.86	2.76	2.7	2.64	2.52	2.46	2.4
15	0.01	3.8	3.67	3.52	3.37	3.29	3.21	3.05	2.96	2.87
15	0.005	4.42	4.25	4.07	3.88	3.79	3.69	3.48	3.37	3.26
20	0.1	1.94	1.89	1.84	1.79	1.77	1.74	1.68	1.64	1.61
20	0.05	2.35	2.28	2.2	2.12	2.08	2.04	1.95	1.9	1.84
20	0.025	2.77	2.68	2.57	2.46	2.41	2.35	2.22	2.16	2.09
20	0.01	3.37	3.23	3.09	2.94	2.86	2.78	2.61	2.52	2.42
20	0.005	3.85	3.68	3.5	3.32	3.22	3.12	2.92	2.81	2.69
24	0.1	1.88	1.83	1.78	1.73	1.7	1.67	1.61	1.57	1.53
24	0.05	2.25	2.18	2.11	2.03	1.98	1.94	1.84	1.79	1.73
24	0.025	2.64	2.54	2.44	2.33	2.27	2.21	2.08	2.01	1.94
24	0.01	3.17	3.03	2.89	2.74	2.66	2.58	2.4	2.31	2.21
24	0.005	3.59	3.42	3.25	3.06	2.97	2.87	2.66	2.55	2.43
30	0.1	1.82	1.77	1.72	1.67	1.64	1.61	1.54	1.5	1.46
30	0.05	2.16	2.09	2.01	1.93	1.89	1.84	1.74	1.68	1.62
30	0.025	2.51	2.41	2.31	2.2	2.14	2.07	1.94	1.87	1.79
30	0.01	2.98	2.84	2.7	2.55	2.47	2.39	2.21	2.11	2.01
30	0.005	3.34	3.18	3.01	2.82	2.73	2.63	2.42	2.3	2.18
60	0.1	1.71	1.66	1.6	1.54	1.51	1.48	1.4	1.35	1.29
60	0.05	1.99	1.92	1.84	1.75	1.7	1.65	1.53	1.47	1.39
60	0.025	2.27	2.17	2.06	1.94	1.88	1.82	1.67	1.58	1.48
60	0.01	2.63	2.5	2.35	2.2	2.12	2.03	1.84	1.73	1.6
60	0.005	2.9	2.74	2.57	2.39	2.29	2.19	1.96	1.83	1.69
120	0.1	1.65	1.6	1.55	1.48	1.45	1.41	1.32	1.26	1.19
120	0.05	1.91	1.83	1.75	1.66	1.61	1.55	1.43	1.35	1.25
120	0.025	2.16	2.05	1.94	1.82	1.76	1.69	1.53	1.43	1.31
120	0.01	2.47	2.34	2.19	2.03	1.95	1.86	1.66	1.53	1.38
120	0.005	2.71	2.54	2.37	2.19	2.09	1.98	1.75	1.61	1.43
Inf	0.1	1.6	1.55	1.49	1.42	1.38	1.34	1.24	1.17	1
Inf	0.05	1.83	1.75	1.67	1.57	1.52	1.46	1.32	1.22	1
Inf	0.025	2.05	1.94	1.83	1.71	1.64	1.57	1.39	1.27	1
Inf	0.01	2.32	2.18	2.04	1.88	1.79	1.7	1.47	1.32	1
Inf	0.005	2.52	2.36	2.19	2	1.9	1.79	1.53	1.36	1

Table A.9: Cumulative binomial probabilities I. Table entry is  $P(X \leq x)$  where  $X \sim \text{bin}(n, p)$ .

Table A.10: Cumulative binomial probabilities II. Table entry is  $P(X \leq x)$  where  $X \sim \text{bin}(n, p)$ .

Table A.11: Cumulative binomial probabilities III. Table entry is  $P(X \leq x)$  where  $X \sim \text{bin}(n, p)$ .

Table A.12: Cumulative binomial probabilities IV. Table entry is  $P(X \leq x)$  where  $X \sim \text{bin}(n, p)$ .

Table A.13: **Cumulative binomial probabilities V.** Table entry is  $P(X \leq x)$  where  $X \sim \text{bin}(n, p)$ .

Table A.14: Cumulative binomial probabilities VI. Table entry is  $P(X \leq x)$  where  $X \sim \text{bin}(n, p)$ .

Table A.15: **Cumulative Poisson probabilities I.** Table entry is  $P(X \leq x)$  where  $X \sim \text{pois}(\lambda)$ .

Table A.16: Cumulative Poisson probabilities II. Table entry is  $P(X \leq x)$  where  $X \sim \text{pois}(\lambda)$ .

Table A.17: **Sign Test I.** Entries give  $P(X \leq x)$  where  $X \sim \text{bin}(n, 1/2)$ . Note that, by symmetry,  $P(X \geq x) = P(X \leq n-x)$ . Cumulative probabilities are give up to the smallest value *greater than or equal to*  $P(X \leq x) = 0.5$

		$n$								
		2	3	4	5	6	7	8	9	10
$x$		0.25	0.125	0.0625	0.0312	0.0156	0.0078	0.0039	0.002	0.001
0		0.75	0.5	0.3125	0.1875	0.1094	0.0625	0.0352	0.0195	0.0107
1		.	.	0.6875	0.5	0.3437	0.2266	0.1445	0.0898	0.0547
2		.	.	.	0.6562	0.5	0.3633	0.2539	0.1719	
3		.	.	.	.	0.6367	0.5	0.377		
4		.	.	.	.	.	0.6367	0.5	0.377	

Table A.18: **Sign Test II.** Entries give  $P(X \leq x)$  where  $X \sim \text{bin}(n, 1/2)$ . Note that, by symmetry,  $P(X \geq x) = P(X \leq n-x)$ . Cumulative probabilities are give up to the smallest value *greater than or equal to*  $P(X \leq x) = 0.5$

		$n$								
		11	12	13	14	15	16	17	18	19
$x$		5e-04	2e-04	1e-04	1e-04	0	0	0	0	0
0		0.0059	0.0032	0.0017	9e-04	5e-04	3e-04	1e-04	1e-04	0
1		0.0327	0.0193	0.0112	0.0065	0.0037	0.0021	0.0012	7e-04	4e-04
2		0.1133	0.073	0.0461	0.0287	0.0176	0.0106	0.0064	0.0038	0.0022
3		0.2744	0.1938	0.1334	0.0898	0.0592	0.0384	0.0245	0.0154	0.0096
4		0.5	0.3872	0.2905	0.212	0.1509	0.1051	0.0717	0.0481	0.0318
5		.	0.6128	0.5	0.3953	0.3036	0.2272	0.1662	0.1189	0.0835
6		.	.	.	0.6047	0.5	0.4018	0.3145	0.2403	0.1796
7		.	.	.	.	0.5982	0.5	0.4073	0.3238	
8		.	.	.	.	.	0.5927	0.5		
9		.	.	.	.	.	.	0.5927	0.5	

Table A.19: **Sign Test III.** Entries give  $P(X \leq x)$  where  $X \sim \text{bin}(n, 1/2)$ . Note that, by symmetry,  $P(X \geq x) = P(X \leq n-x)$ . Cumulative probabilities are give up to the smallest value *greater than or equal to*  $P(X \leq x) = 0.5$

		$n$								
		20	21	22	23	24	25	26	27	28
$x$		0	0	0	0	0	0	0	0	0
0		0	0	0	0	0	0	0	0	0
1		2e-04	1e-04	1e-04	0	0	0	0	0	0
2		0.0013	7e-04	4e-04	2e-04	1e-04	1e-04	0	0	0
3		0.0059	0.0036	0.0022	0.0013	8e-04	5e-04	3e-04	2e-04	1e-04
4		0.0207	0.0133	0.0085	0.0053	0.0033	0.002	0.0012	8e-04	5e-04
5		0.0577	0.0392	0.0262	0.0173	0.0113	0.0073	0.0047	0.003	0.0019
6		0.1316	0.0946	0.0669	0.0466	0.032	0.0216	0.0145	0.0096	0.0063
7		0.2517	0.1917	0.1431	0.105	0.0758	0.0539	0.0378	0.0261	0.0178
8		0.4119	0.3318	0.2617	0.2024	0.1537	0.1148	0.0843	0.061	0.0436
9		0.5881	0.5	0.4159	0.3388	0.2706	0.2122	0.1635	0.1239	0.0925
10		.	.	0.5841	0.5	0.4194	0.345	0.2786	0.221	0.1725
11		.	.	.	0.5806	0.5	0.4225	0.3506	0.2858	
12		.	.	.	.	0.5775	0.5	0.4253		
13		.	.	.	.	.	0.5747			
14		.	.	.	.	.	.			

Table A.20: **Sign Test IV.** Entries give  $P(X \leq x)$  where  $X \sim \text{bin}(n, 1/2)$ . Note that, by symmetry,  $P(X \geq x) = P(X \leq n-x)$ . Cumulative probabilities are given up to the smallest value *greater than or equal to*  $P(X \leq x) = 0.5$

	<i>n</i>								
<i>x</i>	29	30	31	32	33	34	35	36	37
0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0
4	1e-04	0	0	0	0	0	0	0	0
5	3e-04	2e-04	1e-04	1e-04	0	0	0	0	0
6	0.0012	7e-04	4e-04	3e-04	2e-04	1e-04	1e-04	0	0
7	0.0041	0.0026	0.0017	0.0011	7e-04	4e-04	3e-04	2e-04	1e-04
8	0.0121	0.0081	0.0053	0.0035	0.0023	0.0015	9e-04	6e-04	4e-04
9	0.0307	0.0214	0.0147	0.01	0.0068	0.0045	0.003	0.002	0.0013
10	0.068	0.0494	0.0354	0.0251	0.0175	0.0122	0.0083	0.0057	0.0038
11	0.1325	0.1002	0.0748	0.0551	0.0401	0.0288	0.0205	0.0144	0.01
12	0.2291	0.1808	0.1405	0.1077	0.0814	0.0607	0.0448	0.0326	0.0235
13	0.3555	0.2923	0.2366	0.1885	0.1481	0.1147	0.0877	0.0662	0.0494
14	0.5	0.4278	0.3601	0.2983	0.2434	0.1958	0.1553	0.1215	0.0939
15	.	0.5722	0.5	0.43	0.3642	0.3038	0.2498	0.2025	0.162
16	.	.	.	0.57	0.5	0.4321	0.3679	0.3089	0.2557
17	.	.	.	.	.	0.5679	0.5	0.434	0.3714
18	.	.	.	.	.	.	.	0.566	0.5

Table A.21: **Sign Test V.** Entries give  $P(X \leq x)$  where  $X \sim \text{bin}(n, 1/2)$ . Note that, by symmetry,  $P(X \geq x) = P(X \leq n-x)$ . Cumulative probabilities are given up to the smallest value *greater than or equal to*  $P(X \leq x) = 0.5$

Table A.22: **Signed Rank Test I.** Entries give  $P(X \leq x)$  where  $X$  is the Wilcoxon signed rank statistic for sample size  $n$ . Note that, by symmetry,  $P(X \geq x) = P(X \leq n(n+1)/2 - x)$ . Cumulative probabilities are given up to the smallest value *greater than or equal to*  $P(X \leq x) = 0.5$

Table A.23: **Signed Rank Test II.** Entries give  $P(X \leq x)$  where  $X$  is the Wilcoxon signed rank statistic for sample size  $n$ . Note that, by symmetry,  $P(X \geq x) = P(X \leq n(n+1)/2 - x)$ . Cumulative probabilities are give up to the smallest value *greater than or equal to*  $P(X \leq x) = 0.5$

$x$	13	14	15	16	17	18	19	20	21	22	23	24
0	1e-04	1e-04	0	0	0	0	0	0	0	0	0	0
1	2e-04	1e-04	1e-04	0	0	0	0	0	0	0	0	0
2	4e-04	2e-04	1e-04	0	0	0	0	0	0	0	0	0
3	6e-04	3e-04	2e-04	1e-04	0	0	0	0	0	0	0	0
4	9e-04	4e-04	2e-04	1e-04	1e-04	0	0	0	0	0	0	0
5	0.0012	6e-04	3e-04	2e-04	1e-04	0	0	0	0	0	0	0
6	0.0017	9e-04	4e-04	2e-04	1e-04	1e-04	0	0	0	0	0	0
7	0.0023	0.0012	6e-04	3e-04	1e-04	1e-04	0	0	0	0	0	0
8	0.0031	0.0015	8e-04	4e-04	2e-04	1e-04	0	0	0	0	0	0
9	0.004	0.002	0.001	5e-04	3e-04	1e-04	1e-04	0	0	0	0	0
10	0.0052	0.0026	0.0013	7e-04	3e-04	2e-04	1e-04	0	0	0	0	0
11	0.0067	0.0034	0.0017	8e-04	4e-04	2e-04	1e-04	1e-04	0	0	0	0
12	0.0085	0.0043	0.0021	0.0011	5e-04	3e-04	1e-04	1e-04	0	0	0	0
13	0.0107	0.0054	0.0027	0.0013	7e-04	3e-04	2e-04	1e-04	0	0	0	0
14	0.0133	0.0067	0.0034	0.0017	8e-04	4e-04	2e-04	1e-04	1e-04	0	0	0
15	0.0164	0.0083	0.0042	0.0021	0.001	5e-04	3e-04	1e-04	1e-04	0	0	0
16	0.0199	0.0101	0.0051	0.0026	0.0013	6e-04	3e-04	2e-04	1e-04	0	0	0
17	0.0239	0.0123	0.0062	0.0031	0.0016	8e-04	4e-04	2e-04	1e-04	0	0	0
18	0.0287	0.0148	0.0075	0.0038	0.0019	0.001	5e-04	2e-04	1e-04	1e-04	0	0
19	0.0341	0.0176	0.009	0.0046	0.0023	0.0012	6e-04	3e-04	1e-04	1e-04	0	0
20	0.0402	0.0209	0.0108	0.0055	0.0028	0.0014	7e-04	4e-04	2e-04	1e-04	0	0
21	0.0471	0.0247	0.0128	0.0065	0.0033	0.0017	8e-04	4e-04	2e-04	1e-04	1e-04	0
22	0.0549	0.029	0.0151	0.0078	0.004	0.002	0.001	5e-04	3e-04	1e-04	1e-04	0
23	0.0636	0.0338	0.0177	0.0091	0.0047	0.0024	0.0012	6e-04	3e-04	2e-04	1e-04	0
24	0.0732	0.0392	0.0206	0.0107	0.0055	0.0028	0.0014	7e-04	4e-04	2e-04	1e-04	0
25	0.0839	0.0453	0.024	0.0125	0.0064	0.0033	0.0017	8e-04	4e-04	2e-04	1e-04	1e-04
26	0.0955	0.052	0.0277	0.0145	0.0075	0.0038	0.002	0.001	5e-04	3e-04	1e-04	1e-04
27	0.1082	0.0594	0.0319	0.0168	0.0087	0.0045	0.0023	0.0012	6e-04	3e-04	1e-04	1e-04
28	0.1219	0.0676	0.0365	0.0193	0.0101	0.0052	0.0027	0.0014	7e-04	3e-04	2e-04	1e-04
29	0.1367	0.0765	0.0416	0.0222	0.0116	0.006	0.0031	0.0016	8e-04	4e-04	2e-04	1e-04
30	0.1527	0.0863	0.0473	0.0253	0.0133	0.0069	0.0036	0.0018	9e-04	5e-04	2e-04	1e-04
31	0.1698	0.0969	0.0535	0.0288	0.0153	0.008	0.0041	0.0021	0.0011	5e-04	3e-04	1e-04
32	0.1879	0.1083	0.0603	0.0327	0.0174	0.0091	0.0047	0.0024	0.0012	6e-04	3e-04	2e-04
33	0.2072	0.1206	0.0677	0.037	0.0198	0.0104	0.0054	0.0028	0.0014	7e-04	4e-04	2e-04
34	0.2274	0.1338	0.0757	0.0416	0.0224	0.0118	0.0062	0.0032	0.0016	8e-04	4e-04	2e-04
35	0.2487	0.1479	0.0844	0.0467	0.0253	0.0134	0.007	0.0036	0.0019	0.001	5e-04	2e-04
36	0.2709	0.1629	0.0938	0.0523	0.0284	0.0152	0.008	0.0042	0.0021	0.0011	6e-04	3e-04
37	0.2939	0.1788	0.1039	0.0583	0.0319	0.0171	0.009	0.0047	0.0024	0.0013	6e-04	3e-04
38	0.3177	0.1955	0.1147	0.0649	0.0357	0.0192	0.0102	0.0053	0.0028	0.0014	7e-04	4e-04
39	0.3424	0.2131	0.1262	0.0719	0.0398	0.0216	0.0115	0.006	0.0031	0.0016	8e-04	4e-04
40	0.3677	0.2316	0.1384	0.0795	0.0443	0.0241	0.0129	0.0068	0.0036	0.0018	9e-04	5e-04
41	0.3934	0.2508	0.1514	0.0877	0.0492	0.0269	0.0145	0.0077	0.004	0.0021	0.0011	5e-04
42	0.4197	0.2708	0.1651	0.0964	0.0544	0.03	0.0162	0.0086	0.0045	0.0023	0.0012	6e-04
43	0.4463	0.2915	0.1796	0.1057	0.0601	0.0333	0.018	0.0096	0.0051	0.0026	0.0014	7e-04
44	0.473	0.3129	0.1947	0.1156	0.0662	0.0368	0.0201	0.0107	0.0057	0.003	0.0015	8e-04
45	0.5	0.3349	0.2106	0.1261	0.0727	0.0407	0.0223	0.012	0.0063	0.0033	0.0017	9e-04
46	.	0.3574	0.2271	0.1372	0.0797	0.0449	0.0247	0.0133	0.0071	0.0037	0.0019	0.001
47	.	0.3804	0.2444	0.1489	0.0871	0.0494	0.0273	0.0148	0.0079	0.0042	0.0022	0.0011
48	.	0.4039	0.2622	0.1613	0.095	0.0542	0.0301	0.0164	0.0088	0.0046	0.0024	0.0013
49	.	0.4276	0.2807	0.1742	0.1034	0.0594	0.0331	0.0181	0.0097	0.0052	0.0027	0.0014
50	.	0.4516	0.2997	0.1877	0.1123	0.0649	0.0364	0.02	0.0108	0.0057	0.003	0.0016

Table A.24: **Signed Rank Test III.** Entries give  $P(X \leq x)$  where  $X$  is the Wilcoxon signed rank statistic for sample size  $n$ . Note that, by symmetry,  $P(X \geq x) = P(X \leq n(n+1)/2 - x)$ . Cumulative probabilities are give up to the smallest value *greater than or equal to*  $P(X \leq x) = 0.5$

$x$	13	14	15	16	17	18	19	20	21	22	23	24	
51	.	0.4758	0.3193	0.2019	0.1217	0.0708	0.0399	0.022	0.0119	0.0064	0.0034	0.0018	
52	.	0.5	0.3394	0.2166	0.1317	0.077	0.0437	0.0242	0.0132	0.007	0.0037	0.002	
53	.	.	0.3599	0.2319	0.1421	0.0837	0.0478	0.0266	0.0145	0.0078	0.0041	0.0022	
54	.	.	0.3808	0.2477	0.153	0.0907	0.0521	0.0291	0.016	0.0086	0.0046	0.0024	
55	.	.	0.402	0.2641	0.1645	0.0982	0.0567	0.0319	0.0175	0.0095	0.0051	0.0027	
56	.	.	0.4235	0.2809	0.1764	0.1061	0.0616	0.0348	0.0192	0.0104	0.0056	0.0029	
57	.	.	0.4452	0.2983	0.1889	0.1144	0.0668	0.0379	0.021	0.0115	0.0061	0.0033	
58	.	.	0.467	0.3161	0.2019	0.1231	0.0723	0.0413	0.023	0.0126	0.0068	0.0036	
59	.	.	0.489	0.3343	0.2153	0.1323	0.0782	0.0448	0.0251	0.0138	0.0074	0.004	
60	.	.	0.511	0.3529	0.2293	0.1419	0.0844	0.0487	0.0273	0.0151	0.0082	0.0044	
61	.	.	.	0.3718	0.2437	0.1519	0.0909	0.0527	0.0298	0.0164	0.0089	0.0048	
62	.	.	.	0.391	0.2585	0.1624	0.0978	0.057	0.0323	0.0179	0.0098	0.0053	
63	.	.	.	0.4104	0.2738	0.1733	0.1051	0.0615	0.0351	0.0195	0.0107	0.0058	
64	.	.	.	0.4301	0.2895	0.1846	0.1127	0.0664	0.038	0.0212	0.0117	0.0063	
65	.	.	.	0.45	0.3056	0.1964	0.1206	0.0715	0.0411	0.0231	0.0127	0.0069	
66	.	.	.	0.4699	0.3221	0.2086	0.129	0.0768	0.0444	0.025	0.0138	0.0075	
67	.	.	.	0.49	0.3389	0.2211	0.1377	0.0825	0.0479	0.0271	0.015	0.0082	
68	.	.	.	0.51	0.3559	0.2341	0.1467	0.0884	0.0516	0.0293	0.0163	0.0089	
69	.	.	.	.	0.3733	0.2475	0.1562	0.0947	0.0555	0.0317	0.0177	0.0097	
70	.	.	.	.	0.391	0.2613	0.166	0.1012	0.0597	0.0342	0.0192	0.0106	
71	.	.	.	.	0.4088	0.2754	0.1762	0.1081	0.064	0.0369	0.0208	0.0115	
72	.	.	.	.	0.4268	0.2899	0.1868	0.1153	0.0686	0.0397	0.0224	0.0124	
73	.	.	.	.	0.445	0.3047	0.1977	0.1227	0.0735	0.0427	0.0242	0.0135	
74	.	.	.	.	0.4633	0.3198	0.209	0.1305	0.0786	0.0459	0.0261	0.0146	
75	.	.	.	.	0.4816	0.3353	0.2207	0.1387	0.0839	0.0492	0.0281	0.0157	
76	.	.	.	.	.	0.5	0.3509	0.2327	0.1471	0.0895	0.0527	0.0303	0.017
77	.	.	.	.	.	.	0.3669	0.245	0.1559	0.0953	0.0564	0.0325	0.0183
78	.	.	.	.	.	.	0.383	0.2576	0.165	0.1015	0.0603	0.0349	0.0197
79	.	.	.	.	.	.	0.3994	0.2706	0.1744	0.1078	0.0644	0.0374	0.0212
80	.	.	.	.	.	.	0.4159	0.2839	0.1841	0.1145	0.0687	0.0401	0.0228
81	.	.	.	.	.	.	0.4325	0.2974	0.1942	0.1214	0.0733	0.0429	0.0245
82	.	.	.	.	.	.	0.4493	0.3113	0.2045	0.1286	0.078	0.0459	0.0263
83	.	.	.	.	.	.	0.4661	0.3254	0.2152	0.1361	0.0829	0.049	0.0282
84	.	.	.	.	.	.	0.4831	0.3397	0.2262	0.1439	0.0881	0.0523	0.0302
85	.	.	.	.	.	.	0.5	0.3543	0.2375	0.1519	0.0935	0.0557	0.0323
86	.	.	.	.	.	.	.	0.369	0.249	0.1602	0.0991	0.0593	0.0346
87	.	.	.	.	.	.	.	0.384	0.2608	0.1688	0.105	0.0631	0.0369
88	.	.	.	.	.	.	.	0.3991	0.2729	0.1777	0.1111	0.0671	0.0394
89	.	.	.	.	.	.	.	0.4144	0.2853	0.1869	0.1174	0.0712	0.042
90	.	.	.	.	.	.	.	0.4298	0.2979	0.1963	0.1239	0.0755	0.0447
91	.	.	.	.	.	.	.	0.4453	0.3108	0.206	0.1308	0.0801	0.0475
92	.	.	.	.	.	.	.	0.4609	0.3238	0.216	0.1378	0.0848	0.0505
93	.	.	.	.	.	.	.	0.4765	0.3371	0.2262	0.1451	0.0897	0.0537
94	.	.	.	.	.	.	.	0.4922	0.3506	0.2367	0.1527	0.0948	0.057
95	.	.	.	.	.	.	.	0.5078	0.3643	0.2474	0.1604	0.1001	0.0604
96	.	.	.	.	.	.	.	.	0.3781	0.2584	0.1685	0.1056	0.064
97	.	.	.	.	.	.	.	.	0.3921	0.2696	0.1767	0.1113	0.0678
98	.	.	.	.	.	.	.	.	0.4062	0.281	0.1853	0.1172	0.0717
99	.	.	.	.	.	.	.	.	0.4204	0.2927	0.194	0.1234	0.0758
100	.	.	.	.	.	.	.	.	0.4347	0.3046	0.203	0.1297	0.08

Table A.25: **Signed Rank Test IV.** Entries give  $P(X \leq x)$  where  $X$  is the Wilcoxon signed rank statistic for sample size  $n$ . Note that, by symmetry,  $P(X \geq x) = P(X \leq n(n+1)/2 - x)$ . Cumulative probabilities are given up to the smallest value *greater than or equal to*  $P(X \leq x) = 0.5$

<i>x</i>	13	14	15	16	17	18	19	20	21	22	23	24
101	.	.	.	.	.	.	.	0.4492	0.3166	0.2122	0.1363	0.0844
102	.	.	.	.	.	.	.	0.4636	0.3289	0.2217	0.1431	0.089
103	.	.	.	.	.	.	.	0.4782	0.3414	0.2314	0.1501	0.0938
104	.	.	.	.	.	.	.	0.4927	0.354	0.2413	0.1573	0.0987
105	.	.	.	.	.	.	.	0.5073	0.3667	0.2514	0.1647	0.1038
106	.	.	.	.	.	.	.	0.3796	0.2618	0.1723	0.1091	
107	.	.	.	.	.	.	.	0.3927	0.2723	0.1802	0.1146	
108	.	.	.	.	.	.	.	0.4058	0.283	0.1883	0.1203	
109	.	.	.	.	.	.	.	0.4191	0.294	0.1965	0.1261	
110	.	.	.	.	.	.	.	0.4324	0.3051	0.205	0.1322	
111	.	.	.	.	.	.	.	0.4459	0.3164	0.2137	0.1384	
112	.	.	.	.	.	.	.	0.4593	0.3278	0.2226	0.1448	
113	.	.	.	.	.	.	.	0.4729	0.3394	0.2317	0.1514	
114	.	.	.	.	.	.	.	0.4864	0.3512	0.241	0.1583	
115	.	.	.	.	.	.	.	0.5	0.3631	0.2505	0.1653	
116	.	.	.	.	.	.	.	0.5136	0.3751	0.2601	0.1724	
117	.	.	.	.	.	.	.	0.3873	0.27	0.1798		
118	.	.	.	.	.	.	.	0.3995	0.28	0.1874		
119	.	.	.	.	.	.	.	0.4119	0.2902	0.1951		
120	.	.	.	.	.	.	.	0.4243	0.3005	0.2031		
121	.	.	.	.	.	.	.	0.4368	0.311	0.2112		
122	.	.	.	.	.	.	.	0.4494	0.3217	0.2195		
123	.	.	.	.	.	.	.	0.462	0.3325	0.2279		
124	.	.	.	.	.	.	.	0.4746	0.3434	0.2366		
125	.	.	.	.	.	.	.	0.4873	0.3545	0.2454		
126	.	.	.	.	.	.	.	0.5	0.3657	0.2544		
127	.	.	.	.	.	.	.	0.5127	0.377	0.2635		
128	.	.	.	.	.	.	.	.	0.3884	0.2728		
129	.	.	.	.	.	.	.	.	0.3999	0.2823		
130	.	.	.	.	.	.	.	.	0.4115	0.2919		
131	.	.	.	.	.	.	.	.	0.4231	0.3017		
132	.	.	.	.	.	.	.	.	0.4348	0.3115		
133	.	.	.	.	.	.	.	.	0.4466	0.3216		
134	.	.	.	.	.	.	.	.	0.4584	0.3317		
135	.	.	.	.	.	.	.	.	0.4703	0.342		
136	.	.	.	.	.	.	.	.	0.4822	0.3524		
137	.	.	.	.	.	.	.	.	0.4941	0.3629		
138	.	.	.	.	.	.	.	.	0.5059	0.3735		
139	.	.	.	.	.	.	.	.	.	0.3841		
140	.	.	.	.	.	.	.	.	.	0.3949		
141	.	.	.	.	.	.	.	.	.	0.4058		
142	.	.	.	.	.	.	.	.	.	0.4167		
143	.	.	.	.	.	.	.	.	.	0.4277		
144	.	.	.	.	.	.	.	.	.	0.4387		
145	.	.	.	.	.	.	.	.	.	0.4498		
146	.	.	.	.	.	.	.	.	.	0.4609		
147	.	.	.	.	.	.	.	.	.	0.4721		
148	.	.	.	.	.	.	.	.	.	0.4832		
149	.	.	.	.	.	.	.	.	.	0.4944		
150	.	.	.	.	.	.	.	.	.	0.5056		

Table A.26: **Rank Sum Test I.** Entries give  $P(X \leq x)$  where  $X$  is the Wilcoxon rank sum statistic for sample sizes  $n_1 \leq n_2$ . Note that, by symmetry,  $P(X \geq x) = P(X \leq n_1(n_1+n_2+1)/2 - x)$ . Cumulative probabilities are given up to the smallest value *greater than or equal to*  $P(X \leq x) = 0.5$

Table A.27: **Rank Sum Test II.** Entries give  $P(X \leq x)$  where  $X$  is the Wilcoxon rank sum statistic for sample sizes  $n_1 \leq n_2$ . Note that, by symmetry,  $P(X \geq x) = P(X \leq n_1(n_1+n_2+1)/2 - x)$ . Cumulative probabilities are given up to the smallest value *greater than or equal to*  $P(X \leq x) = 0.5$

Table A.28: **Rank Sum Test III.** Entries give  $P(X \leq x)$  where  $X$  is the Wilcoxon rank sum statistic for sample sizes  $n_1 \leq n_2$ . Note that, by symmetry,  $P(X \geq x) = P(X \leq n_1(n_1+n_2+1)/2 - x)$ . Cumulative probabilities are given up to the smallest value *greater than or equal to*  $P(X \leq x) = 0.5$

Table A.29: **Rank Sum Test IV.** Entries give  $P(X \leq x)$  where  $X$  is the Wilcoxon rank sum statistic for sample sizes  $n_1 \leq n_2$ . Note that, by symmetry,  $P(X \geq x) = P(X \leq n_1(n_1+n_2+1)/2 - x)$ . Cumulative probabilities are given up to the smallest value *greater than or equal to*  $P(X \leq x) = 0.5$

$n_1 =$	5	5	5	5	5	5
$n_2 =$	5	6	7	8	9	10
$x =$	15	0.004	0.0022	0.0013	8e-04	5e-04
	16	0.0079	0.0043	0.0025	0.0016	0.001
	17	0.0159	0.0087	0.0051	0.0031	0.002
	18	0.0278	0.0152	0.0088	0.0054	0.0035
	19	0.0476	0.026	0.0152	0.0093	0.006
	20	0.0754	0.0411	0.024	0.0148	0.0095
	21	0.1111	0.0628	0.0366	0.0225	0.0145
	22	0.1548	0.0887	0.053	0.0326	0.021
	23	0.2103	0.1234	0.0745	0.0466	0.03
	24	0.2738	0.1645	0.101	0.0637	0.0415
	25	0.3452	0.2143	0.1338	0.0855	0.0559
	26	0.4206	0.2684	0.1717	0.1111	0.0734
	27	0.5	0.3312	0.2159	0.1422	0.0949
	28	.	0.3961	0.2652	0.1772	0.1199
	29	.	0.4654	0.3194	0.2176	0.1489
	30	.	0.5346	0.3775	0.2618	0.1818
	31	.	.	0.4381	0.3108	0.2188
	32	.	.	0.5	0.3621	0.2592
	33	.	.	.	0.4165	0.3032
	34	.	.	.	0.4716	0.3497
	35	.	.	.	0.5284	0.3986
	36	.	.	.	.	0.4491
	37	.	.	.	.	0.5
	38	.	.	.	.	.
	39	.	.	.	.	.
	40	.	.	.	.	.

Table A.30: **Rank Sum Test V**. Entries give  $P(X \leq x)$  where  $X$  is the Wilcoxon rank sum statistic for sample sizes  $n_1 \leq n_2$ . Note that, by symmetry,  $P(X \geq x) = P(X \leq n_1(n_1+n_2+1)/2 - x)$ . Cumulative probabilities are given up to the smallest value *greater than or equal to*  $P(X \leq x) = 0.5$

Table A.31: **Rank Sum Test VI.** Entries give  $P(X \leq x)$  where  $X$  is the Wilcoxon rank sum statistic for sample sizes  $n_1 \leq n_2$ . Note that, by symmetry,  $P(X \geq x) = P(X \leq n_1(n_1+n_2+1)/2 - x)$ . Cumulative probabilities are given up to the smallest value *greater than or equal to*  $P(X \leq x) = 0.5$

$n_1 =$	8	8	8	9	9	10
$n_2 =$	8	9	10	9	10	10
$x =$	36	1e-04	0	0	0	0
	37	2e-04	1e-04	0	0	0
	38	3e-04	2e-04	1e-04	0	0
	39	5e-04	3e-04	2e-04	0	0
	40	9e-04	5e-04	3e-04	0	0
	41	0.0015	8e-04	4e-04	0	0
	42	0.0023	0.0012	7e-04	0	0
	43	0.0035	0.0019	0.001	0	0
	44	0.0052	0.0028	0.0015	0	0
	45	0.0074	0.0039	0.0022	0	0
	46	0.0103	0.0056	0.0031	0	0
	47	0.0141	0.0076	0.0043	1e-04	0
	48	0.019	0.0103	0.0058	1e-04	1e-04
	49	0.0249	0.0137	0.0078	2e-04	1e-04
	50	0.0325	0.018	0.0103	4e-04	2e-04
	51	0.0415	0.0232	0.0133	6e-04	3e-04
	52	0.0524	0.0296	0.0171	9e-04	5e-04
	53	0.0652	0.0372	0.0217	0.0014	7e-04
	54	0.0803	0.0464	0.0273	0.002	0.0011
	55	0.0974	0.057	0.0338	0.0028	0.0015
	56	0.1172	0.0694	0.0416	0.0039	0.0021
	57	0.1393	0.0836	0.0506	0.0053	0.0028
	58	0.1641	0.0998	0.061	0.0071	0.0038
	59	0.1911	0.1179	0.0729	0.0094	0.0051
	60	0.2209	0.1383	0.0864	0.0122	0.0066
	61	0.2527	0.1606	0.1015	0.0157	0.0086
	62	0.2869	0.1852	0.1185	0.02	0.011
	63	0.3227	0.2117	0.1371	0.0252	0.014
	64	0.3605	0.2404	0.1577	0.0313	0.0175
	65	0.3992	0.2707	0.18	0.0385	0.0217
	66	0.4392	0.3029	0.2041	0.047	0.0267
	67	0.4796	0.3365	0.2299	0.0567	0.0326
	68	0.5204	0.3715	0.2574	0.068	0.0394
	69	.	0.4074	0.2863	0.0807	0.0474
	70	.	0.4442	0.3167	0.0951	0.0564
	71	.	0.4813	0.3482	0.1112	0.0667
	72	.	0.5187	0.3809	0.129	0.0782
	73	.	.	0.4143	0.1487	0.0912
	74	.	.	0.4484	0.1701	0.1055
	75	.	.	0.4827	0.1933	0.1214
	76	.	.	0.5173	0.2181	0.1388
	77	.	.	.	0.2447	0.1577
	78	.	.	.	0.2729	0.1781
	79	.	.	.	0.3024	0.2001
	80	.	.	.	0.3332	0.2235
	81	.	.	.	0.3652	0.2483
	82	.	.	.	0.3981	0.2745
	83	.	.	.	0.4317	0.3019
	84	.	.	.	0.4657	0.3304
	85	.	.	.	0.5	0.3598
						0.0716

Table A.32: **Rank Sum Test VII.** Entries give  $P(X \leq x)$  where  $X$  is the Wilcoxon rank sum statistic for sample sizes  $n_1 \leq n_2$ . Note that, by symmetry,  $P(X \geq x) = P(X \leq n_1(n_1+n_2+1)/2 - x)$ . Cumulative probabilities are given up to the smallest value *greater than or equal to*  $P(X \leq x) = 0.5$

$n_1 =$	8	8	8	9	9	10
$n_2 =$	8	9	10	9	10	10
$x =$	86	.	.	0.5343	0.3901	0.0827
	87	.	.	.	0.4211	0.0952
	88	.	.	.	0.4524	0.1088
	89	.	.	.	0.4841	0.1237
	90	.	.	.	0.5159	0.1399
	91	.	.	.	.	0.1575
	92	.	.	.	.	0.1763
	93	.	.	.	.	0.1965
	94	.	.	.	.	0.2179
	95	.	.	.	.	0.2406
	96	.	.	.	.	0.2644
	97	.	.	.	.	0.2894
	98	.	.	.	.	0.3153
	99	.	.	.	.	0.3421
	100	.	.	.	.	0.3697
	101	.	.	.	.	0.398
	102	.	.	.	.	0.4267
	103	.	.	.	.	0.4559
	104	.	.	.	.	0.4853
	105	.	.	.	.	0.5147

# Appendix B

# Mathematical Review

## B.1 Conventions and Notation

The set of (finite) real numbers is denoted  $\mathbb{R}$ , and the set of extended real numbers is denoted  $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$ . The restriction to nonnegative real numbers is written  $\mathbb{R}_{>0} = [0, \infty)$  and  $\bar{\mathbb{R}}_{>0} = \mathbb{R}_{>0} \cup \{\infty\}$ . We use standard notation for open, closed, left closed and right closed intervals  $(a, b)$ ,  $[a, b]$ ,  $[a, b)$ ,  $(a, b]$ . A reference to a interval  $I$  on  $\bar{\mathbb{R}}$  may be any of these types.

The set of (finite) integers will be denoted  $\mathbb{I}$ , while the extended integers will be  $\bar{\mathbb{I}} = \mathbb{I} \cup \{-\infty, \infty\}$ . The set of natural numbers  $\mathbb{N}$  is taken to be the set of positive integers, which  $\mathbb{N}_0$  is the set of nonnegative integers. A rational number is any real number expressible as a ratio of integers.

The absolute value of  $a \in \mathbb{R}$  is denoted  $|a| = \sqrt{a^2}$ , while  $|z| = (z\bar{z})^{1/2} = (a^2 + b^2)^{1/2} \in \mathbb{R}$  is also known as the magnitude or modulus of  $z \in \mathbb{C}$ .

If  $\mathcal{S}$  is a set of any type of number,  $\mathcal{S}^d$ ,  $d \in \mathbb{N}$ , denotes the set of  $d$ -dimensional vectors  $\tilde{s} = (s_1, \dots, s_d)$ , which are ordered collections of numbers  $s_i \in \mathcal{S}$ . In particular, the set of  $d$ -dimensional real vectors is written  $\mathbb{R}^d$ . When  $0, 1 \in \mathcal{S}$ , we may write the zero or one vector  $\vec{0} = (0, \dots, 0)$ ,  $\vec{1} = (1, \dots, 1)$ , so that  $c\vec{1} = (c, \dots, c)$ .

A collection of  $d$  numbers from  $\mathcal{S}$  is *unordered* if no reference is made to the order (they are unlabelled). Otherwise the collection is *ordered*, that is, it is a vector. An unordered collection from  $\mathcal{S}$  differs from a set in that a number  $s \in \mathcal{S}$  may be represented more than once. Braces  $\{\dots\}$  enclose a set while parentheses  $(\dots)$  enclose a vector (braces will also be used to denote indexed sequences, when the context is clear).

The following notation will be used

$\in$	is an element of
$\notin$	is not an element of
$\subset$	is a subset of
$\exists$	there exists
$\ni$	such that
$\Rightarrow$	implies
$\iff$	if and only if

Sets will often be represented in the following general form:

$$A = \{ \text{type of object } x : \text{property } x \text{ must possess in order to be included in } A \}.$$

The set of points in the Cartesian plain for which the components are within one unit of each other is written:

$$A = \{(x, y) \in \mathbb{R}^2 : |x - y| \leq 1\}.$$

The *empty set* (the set containing no elements) is a well defined set and is denoted  $\emptyset$ .

A *function*  $f$  maps a *domain*  $\mathcal{X}$  to a *range* (or *codomain*)  $\mathcal{Y}$ , which is denoted  $f : \mathcal{X} \mapsto \mathcal{Y}$ . This means  $f$  associates exactly one element of  $\mathcal{Y}$  to all elements of  $\mathcal{X}$ . If  $\mathcal{Y} \subset \mathbb{R}$ , we say  $f$  is a *real valued function*, or simply *real function*. Usually, when we refer to a function, it can be assumed that  $\mathcal{X} \subset \mathbb{R}$  and  $\mathcal{Y} \subset \mathbb{R}$ .

Note that it is not necessary that all elements  $y$  of the range  $\mathcal{Y}$  be represented by the function  $f$ , in the sense that there exists at least one  $x \in \mathcal{X}$  for which  $y = f(x)$ . The subset of  $\mathcal{Y}$  that is so represented is called the *image*:

$$\text{image} = f(\mathcal{X}) = \{y \in \mathcal{Y} : \exists x \in \mathcal{S} \ni f(x) = y\}.$$

A function  $f : \mathcal{X} \mapsto \mathcal{Y}$  is called *one-to-one* or *injective* if  $f(x) \neq f(x')$  for all distinct elements  $x, x'$  of  $\mathcal{X}$ . In addition,  $f$  is *onto* or *surjective* if the image is the same as the range (for all  $y \in \mathcal{Y} \exists x \in \mathcal{X} \ni f(x) = y$ ). If  $f$  is both onto and one-to-one, it is a *bijective* function. A bijective function induces a *one-to-one correspondence* between the domain and the range.

Note that a one-to-one function becomes a bijective function when the range is replaced by the image. When we say that  $f$  is a one-to-one function, we are generally interested in the one-to-one correspondence between the

domain and the image, even when the image is strictly smaller than the range.

A bijective function  $f : \mathcal{X} \mapsto \mathcal{Y}$  has an inverse function  $f^{-1} : \mathcal{Y} \mapsto \mathcal{X}$  for which  $f^{-1}(y) = x$  if and only if  $f(x) = y$ . Formally, a function must be bijective to possess an inverse (be *invertible*). However, we may informally refer to the inverse of a one-to-one (but not necessarily bijective) function, if it is understood that the range has been replaced by the image. We always have  $f^{-1}(f(x)) = x$  and  $f(f^{-1}(y)) = y$ .

Given function  $f : \mathcal{X} \mapsto \mathcal{Y}$  we may refer to the image of a subset  $E \subset \mathcal{X}$  of the domain as the values attained by  $f$  for all  $x \in E$ , which is written

$$f(E) = \{f(x) : x \in E\}.$$

Similarly the *preimage* or *inverse image* of a subset  $F \in \mathcal{Y}$  of the range is the set of all  $x \in \mathcal{X}$  for which  $f(x) \in F$ , which is written

$$f^{-1}(F) = \{x \in \mathcal{X} : f(x) \in F\}.$$

Note that a preimage may be evaluated even if  $f$  is not invertible, despite the notation. In fact, we may have  $f^{-1}F = \emptyset$ .

## B.2 Infimum, Supremum, Minimum and Maximum

Suppose  $E \in \mathbb{R}$ . Then the *infimum* and *supremum* of  $E$  are, respectively,

$$\begin{aligned}\inf E &= \text{the greatest lower bound (GLB) of } E, \text{ and} \\ \sup E &= \text{the least upper bound (LUB) of } E.\end{aligned}$$

That is,  $\inf E = x$  if  $x \leq y$  for all  $y \in E$ , and is the largest number with this property. Similarly,  $\sup E = x$  if  $x \geq y$  for all  $y \in E$ , and is the smallest number with this property.

The infimum and supremum are always defined, even for the empty set:

$$\begin{aligned}\inf \emptyset &= \infty, \text{ and} \\ \sup \emptyset &= -\infty.\end{aligned}$$

The *minimum* and *maximum* of  $E$  are similar to the infimum and supremum, and defined as

$$\begin{aligned}\min E &= \text{the smallest number in } E, \text{ and} \\ \max E &= \text{the largest number in } E,\end{aligned}$$

if such numbers exist, and they might not. In contrast,  $\inf E$  and  $\sup E$  need not be elements of  $E$ , and always exist. If the minimum and maximum do exist, then

$$\begin{aligned}\min E &= \inf E, \text{ and} \\ \max E &= \sup E.\end{aligned}$$

**Example B.1** We have

$$\begin{aligned}\inf [0, 1] &= 0, \\ \sup [0, 1] &= 1, \\ \min [0, 1] &= 0, \\ \max [0, 1] &= 1,\end{aligned}$$

but

$$\begin{aligned}\inf (0, 1) &= 0, \\ \sup (0, 1) &= 1, \\ \min (0, 1) &\quad \text{does not exist,} \\ \max (0, 1) &\quad \text{does not exist.}\end{aligned}$$

■

### B.3 Limits

A sequence of real numbers  $x_1, x_2, \dots$  converges to limit  $y$ , written

$$\lim_{i \rightarrow \infty} x_i = y$$

if for all  $\epsilon > 0$  there exists index  $n_\epsilon$  such that  $|x_n - y| < \epsilon$  for all  $n \geq n_\epsilon$ .

### B.4 Matrices

Let  $M_{m,n}$  be the set of  $m \times n$  matrices  $A$ , for which  $A_{i,j} \in \mathbb{R}$  (or, when required for clarity,  $[A]_{i,j} \in \mathbb{K}$ ) is the element of the  $i$ th row and  $j$ th column. The *square matrices* are denoted as  $M_m = M_{m,m}$ . Elements of  $M_{m,1}$  are *column vectors* and elements of  $M_{1,m}$  are *row vectors*. A matrix in  $M_{m,n}$  is equivalently an ordered set of  $m$  row vectors or  $n$  column vectors. The transpose  $A^T \in M_{n,m}$  of a matrix  $A \in M_{m,n}$  has elements  $A'_{j,i} = A_{i,j}$ . For  $A \in M_{n,k}$ ,  $B \in M_{k,m}$  we always understand matrix multiplication to mean that  $C = AB$  possesses elements  $C_{i,j} = \sum_{k'=1}^k A_{i,k'} B_{k',j}$ , so that

matrix multiplication is generally not commutative. Then  $(A^T)^T = A$  and  $(AB)^T = B^T A^T$  where the product is permitted.

A matrix  $A \in M_n$  is *diagonal* if the only nonzero elements are on the diagonal, and can therefore be referred to by the diagonal elements  $\text{diag}(a_1, \dots, a_n) = \text{diag}(A_{1,1}, \dots, A_{n,n})$ . A diagonal matrix is *positive diagonal* or *nonnegative diagonal* if all diagonal elements are positive or nonnegative.

The identity matrix  $I \in M_m$  is the matrix uniquely possessing the property that  $A = IA = AI$  for all  $A \in M_m$ . For any matrix  $A \in M_m$  there exists at most one matrix  $A^{-1} \in M_m$  for which  $AA^{-1} = I$ , referred to as the *inverse* of  $A$ . An inverse need not exist (for example, if the elements of  $A$  are constant).

## B.5 Geometric Series

We will make use of the following geometric series:

$$\begin{aligned} \sum_{i=0}^{\infty} \frac{(i+m)!}{i!} r^i &= \frac{m!}{(1-r)^{m+1}} \text{ for } r^2 < 1, \quad m = 0, 1, 2, \dots \\ \sum_{i=0}^n r^i &= \frac{1 - r^{n+1}}{1 - r} \text{ for } r \neq 1. \end{aligned} \quad (\text{B.1})$$

This gives special cases:

$$\begin{aligned} \sum_{i=0}^{\infty} r^i &= r^0 + r^1 + r^2 + \dots = \frac{1}{1-r}, \text{ for } m = 0 \\ \sum_{i=0}^{\infty} (i+1)r^i &= \sum_{i=1}^{\infty} ir^{i-1} = r^0 + 2r^1 + 3r^2 + \dots = \frac{1}{(1-r)^2}, \text{ for } m = 1 \\ \sum_{i=0}^{\infty} (i+2)(i+1)r^i &= \sum_{i=1}^{\infty} (i+1)ir^{i-1} = 2 \times 1 \times r^0 + 3 \times 2 \times r^1 + 4 \times 3 \times r^2 + \dots \\ &= \frac{2}{(1-r)^3}, \text{ for } m = 2. \end{aligned}$$

## B.6 Binomial Theorem

Let  $a, b$  be any two numbers, and let  $n$  be a positive integer. Then

$$(a+b)^n = \sum_{i=1}^n a^i b^{n-i} \binom{n}{i} \quad (\text{B.2})$$

## B.7 Gamma Function

The gamma function is defined by the indefinite integral

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx. \quad (\text{B.3})$$

It is defined for all real numbers  $t$ , except for 0 and all negative integers. Our interest will be in values  $t > 0$ . By a change of variable argument, we have the following integral, given constant  $a > 0$ :

$$\int_0^\infty x^{t-1} e^{-ax} = a^{-t} \Gamma(t). \quad (\text{B.4})$$

The gamma function is related to the factorial  $n!$  (see Definition 3.3). For integers  $n = 1, 2, \dots$

$$\Gamma(n) = (n-1)! = (n-1) \times (n-2) \times \dots \times 1.$$

## Appendix C

# Introduction to R

R is an interpretive programming environment designed primarily for statistical computations, but which extends into more general applications of optimization and numerical analysis. An extensive library of statistical algorithms is accessible from within R and through curated software repositories such as [cran.r-project.org](http://cran.r-project.org), [bioconductor.org](http://bioconductor.org) and [omegahat.org](http://omegahat.org). To a large degree, the operation of R is independent of the operating system, but sometimes OS specific issues arise.

R is based on a command line interpreter. Its commands can also be used to construct high level programs permitting standard looping and conditional execution statements. It has especially powerful graphics abilities.

### C.1 Mathematical Operations on Scalars and Vectors

A good first command is

```
> help(Arithmetc)
```

This produces documentation on the use of arithmetic binary operators. The format of the documentation will depend on the operating system.

#### Description

These binary operators perform arithmetic on numeric or complex vectors (or objects which can be coerced to them).

#### Usage

```
x + y  
x - y
```

```
x * y
x / y
x ^ y
x %% y
x %/% y
Arguments
```

`x, y`  
 numeric or complex vectors or objects which can be coerced to such,  
 or other objects for which methods have been written.

<remaining documentation omitted>

To make use of these operators we can assign numbers to variables `x` and `y`:

```
> x = 13
> y = 4
> x + y
[1] 17
> x - y
[1] 9
> x * y
[1] 52
> x / y
[1] 3.25
> x ^ y
[1] 28561
> x %% y
[1] 1
> x %/% y
[1] 3
```

Comments are defined by the `#` symbol

```
> # this is a comment
> x = 3
> # x = 2
> x
[1] 3
>
```

Assignment can also be done with the syntax

```
> x <- 13
```

```
> y <- 4
> x
[1] 13
> y
[1] 4
>
```

Note that R is case sensitive:

```
> a = 4
> A = 3
> a - A
[1] 1
> help(arithmetic)
No documentation for arithmetic in specified packages and libraries:
you could try ??arithmetic
>
```

### C.1.1 Vectors in R

In the above examples, an index reference [1] is given whenever a variable is displayed (simply by typing the name of the variable). This is because R considers a single number as a special case of a vector (or array). One of the advantages of R for statistical computation is the manner in which vectors can be manipulated almost as easily as variables. Vectors are easily constructed using the `c()` function (you can always enter `help(c)`):

```
> x = c(3.4, -1, 99, 1/2)
> x
[1] 3.4 -1.0 99.0 0.5
```

An element of a vector is referenced using square brackets:

```
> x = c(3.4, -1, 99, 1/2)
> x[2]
[1] -1
> i = 3
> x[i]
[1] 99
```

Note that the `c()` function accepts mathematical expressions as argument (such as  $1/2$  or  $\sin(0.075)$ ), which are evaluated when the vector object is created.

Scalar operations on vectors follow intuitive rules:

```
> x = c(1, 3, 10, 99.3)
> x
[1] 1.0 3.0 10.0 99.3
> x+1
[1] 2.0 4.0 11.0 100.3
> x*10
[1] 10 30 100 993
```

Addition and multiplication of two vectors of equal length is done by element:

```
> x = c(1, 3, 10, 99.3)
> y = c(1, 1, 2, 2)
> x*y
[1] 1.0 3.0 20.0 198.6
> x + y
[1] 2.0 4.0 12.0 101.3
> x + 2*y
[1] 3.0 5.0 14.0 103.3
> x + 2*y*x*x
[1] 3.00 21.00 410.00 39541.26
```

It is important to note that R will *attempt* to add, subtract, multiply or divide vectors of unequal length. The decision made by R on your behalf is to recycle the shorter length vector enough times to permit the operation. If the length of the longer vector is not a multiple of the length of the shorter vector, R will give a warning:

```
> c(1,2,3) + c(10,20)
[1] 11 22 13
Warning message:
In c(1, 2, 3) + c(10, 20) :
  longer object length is not a multiple of shorter object length
```

In the above example, the shorter vector `c(10,20)` was extended to `c(10,20,10)` for the purposes of evaluation. In the following example, the shorter vector `c(10,20)` is extended to `c(10,20,10,20)`. In this case R will give no warning:

```
> c(1,2,3,4) + c(10,20)
[1] 11 22 13 24
```

The following examples should give an idea how this works. Note that only the last expression produces a warning.

```
> c(1,2,1,2,1,2,1,2) + c(10,20)
```

```
[1] 11 22 11 22 11 22 11 22
> c(100,200) + c(1,2,1,2,1,2,1,2)
[1] 101 202 101 202 101 202 101 202
> c(100,200)*c(1,2,1,2,1,2,1,2)
[1] 100 400 100 400 100 400 100 400
> c(50,75)/c(1,2,1,2,1,2,1,2)
[1] 50.0 37.5 50.0 37.5 50.0 37.5 50.0 37.5
> c(100,200) - c(1,2,1,2,1,2,1,2)
[1] 99 198 99 198 99 198 99 198
> c(100,200) + c(1,2,1,2,1,2,1,2,1)
[1] 101 202 101 202 101 202 101 202 101
Warning message:
In c(100, 200) + c(1, 2, 1, 2, 1, 2, 1, 2, 1) :
  longer object length is not a multiple of shorter object length
```

Special vectors can be created in a number of ways:

```
> 1:5
[1] 1 2 3 4 5
> 5:1
[1] 5 4 3 2 1
> 1:5 + 5:1
[1] 6 6 6 6 6
> rep(3,10)
[1] 3 3 3 3 3 3 3 3 3 3
> rep(3:1,4)
[1] 3 2 1 3 2 1 3 2 1 3 2 1
> seq(0, 1, 0.25)
[1] 0.00 0.25 0.50 0.75 1.00
> seq(0, 1, 0.27)
[1] 0.00 0.27 0.54 0.81
```

The length of a vector is given by the function `length()`:

```
> x = rep(3:1,4)
> length(x)
[1] 12
```

Vectors of length 0 are defined in R and are assigned the symbol `NULL`

```
> x = NULL
> length(x)
[1] 0
> x = null
Error: object 'null' not found
```

R also has a missing value indicator `NA` which may be used in place of a number (formally, an element of a vector):

```
> c(1,NA,3)
[1] 1 NA 3
> c(1,NA,3)+2
[1] 3 NA 5
```

Conventions for dealing with missing values will depend on the particular function. For example, we can input a vector into the function `sum()`:

```
> x = c(2,4,6)
> sum(x)
[1] 12
```

If the vector has a missing value we get the following result:

```
> x = c(2,NA,6)
> sum(x)
[1] NA
```

unless we specify that missing values are to be removed by setting the remove missing value option `na.rm` to `TRUE`:

```
> x = c(2,NA,6)
> sum(x, na.rm=TRUE)
[1] 8
> sum(x, na.rm=T)
[1] 8
> sum(x, na.rm=TR)
Error: object 'TR' not found
```

If we consult the `sum()` documentation by using `help(sum)` we find that `na.rm = FALSE` is the default option.

R has some flexibility regarding undefined numbers, for example:

```
> 1/0
[1] Inf
> 1/c(0,0,0)
[1] Inf Inf Inf
> x = 1/c(0,0,0)
> x
[1] Inf Inf Inf
```

Note that R offers considerable flexibility when appending elements to vectors. A reference to an element beyond the length of the vector yields `NA`. However, an assignment to an element beyond the length of the vector forces an extension of the length of the vector.

```
> x = c(5,4,3,2,1)
> x
[1] 5 4 3 2 1
> x[10]
[1] NA
> x[10] = 999
> x
[1] 5 4 3 2 1 NA NA NA NA 999
```

### C.1.2 Global Options

Many system options can be changed. The function `options()` can be used to display the value of an option by inputting a character string holding the option name. For example, the character used as the R command line is an option with the name ‘prompt’:

```
> options("prompt")
$prompt
[1] ">"
```

We can also change options within the `option()` function argument. For example, we can include an inspirational message in the prompt:

```
> options(prompt = "Do It Right the First Time> ")
Do It Right the First Time> options('prompt')
$prompt
[1] "Do It Right the First Time> "
```

If we eventually change our mind, we can always return to the original prompt:

```
Do It Right the First Time> options(prompt = "> ")
> options('prompt')
$prompt
[1] "> "

> "That's better ..."
[1] "That's better ..."
```

This should be done sparingly, but, occasionally, a contributed function may change an option, so it’s good to be aware of this possibility.

To see all options enter the `option()` command without argument:

```
> options()
$add.smooth
```

```
[1] TRUE

$bitmapType
[1] "quartz"

$browser
[1] "/usr/bin/open"

$browserNLdisabled
[1] FALSE

$CBoundsCheck
[1] FALSE

$check.bounds
[1] FALSE

$citation.bibtex.max
[1] 1

$continue
[1] "+"

<remaining output omitted>
```

### C.1.3 Modes (or Types)

An element of a vector is stored as one of several types (in R the term *mode* is also used). Numerical data can be stored as an *integer* or a *double* type (that is, integer or real). Usually, operations in R are not highly dependent on the difference between integer and real.

```
> x = 4:6
> x[2]
[1] 5
> x[2.2]
[1] 5
> x[2.5]
[1] 5
> x[2.9]
[1] 5
> x[2.99999]
```

```
[1] 5
> x[3]
[1] 6
```

The distinction, however, is important when R passes data to other programs written in, for example, C++.

It is important to realize that the number stored is not necessarily the number displayed:

```
> x = 1.2344556543
> x
[1] 1.234456
> y = x - 1.234456
> y == 0
[1] FALSE
> y
[1] -3.457e-07
```

There is a default rounding of 7 significant digits for R displays.

```
> options('digits')
$digits
[1] 7

> 2.3433345994
[1] 2.343335
> options(digits=3)
> 2.3433345994
[1] 2.34
> options(digits=7)
> 2.3433345994
[1] 2.343335
> options('digits')
$digits
[1] 7
```

Types can be *coerced* into other types. Real numbers are rounded down when this happens.

```
> 3 == 2.99999999
[1] FALSE
> as.integer(2.99999999)
[1] 2
> as.integer(3.00000001)
[1] 3
```

Real numbers can be rounded off (`round()`), `signif()`), or converted to integers with the `floor()` or `ceiling()` functions, or with the `round()` function with option `digits=0`:

```
> x = 34.59996
> round(x,3)
[1] 34.6
> round(x,5)
[1] 34.59996
> round(x,0)
[1] 35
> ceiling(x)
[1] 35
> floor(x)
[1] 34
> round(0.003344,3)
[1] 0.003
> signif(0.003344,3)
[1] 0.00334
```

Complex numbers are also supported (see `help(complex)`).

Character strings can be stored in variables or vectors:

```
> days = c("Monday", "Tuesday", "Wednesday", "Thursday",
  "Friday", "Saturday", "Sunday")
> days
[1] "Monday"      "Tuesday"      "Wednesday"   "Thursday"
[5] "Friday"       "Saturday"    "Sunday"
```

Vectors can also be of *logical* type, taking values `TRUE` and `FALSE` (or `T` and `F`). These can be entered directly, or be produced as the value of a *conditional expression*:

```
> 4 > 1
[1] TRUE
> 4 < 1
[1] FALSE
> 4 == 6
[1] FALSE
> 4 > 4
[1] FALSE
> 4 >= 4
[1] TRUE
> 4 <= 4
```

```
[1] TRUE
> 4 != 6
[1] TRUE
```

Note that `==` is the logical relationship ‘is equal to’, while `=` is the assign operator. Confusing the two can lead to bugs which might be hard to detect. Also, `!=` means ‘is not equal to’. As expected `<`, `>`, `<=`, `>=` mean ‘less than’, ‘greater than’, ‘less than or equal to’, ‘greater than or equal to’.

```
> x = c(2<2, 2>2, 2<=2, 2>=2, 2==2, 2!=2)
> x
[1] FALSE FALSE  TRUE  TRUE  TRUE FALSE
```

The logical (or Boolean) operators are `!` (negation); `&` or `&&` (logical AND); `|` or `||` (logical OR), `xor()` (exclusive OR). The difference between `&` and `&&` is that `&` yields elementwise evaluation of vectors, while `&&` is intended for single Boolean values. It should be noted that `&&` can be applied to vectors, but will examine elements left to right, using only the first. Confusing the two forms can lead to unexpected results. The same comment applies to `|` and `||`. The function `xor()` accepts two logical values as arguments, returning `TRUE` if and only one of the arguments equals `TRUE`. This function can do elementwise evaluation of vectors:

```
> x = T
> y = F
> x & y
[1] FALSE
> x && y
[1] FALSE
> x | y
[1] TRUE
> x || y
[1] TRUE
> x == y
[1] FALSE
> x != y
[1] TRUE
> xor(x,y)
[1] TRUE
>
> # try vectors
> x = c(T,F)
> y = c(F,F)
> x & y
```

```
[1] FALSE FALSE
> x && y
[1] FALSE
> x | y
[1] TRUE FALSE
> x || y
[1] TRUE
> xor(x,y)
[1] TRUE FALSE
```

Note that logical vectors can be constructed by applying conditional statements to other vectors:

```
> x = seq(0, 50, 7)
> x
[1] 0 7 14 21 28 35 42 49
> z = (x<25)
> z
[1] TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE
> z = (x %% 3 == 0)
> z
[1] TRUE FALSE FALSE TRUE FALSE FALSE TRUE FALSE
```

In the preceding example, we identify all elements of a vector `x` which are less than 25, and which are divisible by 3.

Modes have specific functions associated with them to create objects and test for object type:

```
# 
# Make integer object
#
x = integer()
x
length(x)
#
# Give it length 1
#
x = integer(1)
x
length(x)
#
# Create an integer vector of length 5
#
x = integer(5)
```

```
x
length(x)
#
# Verify that it is an integer object
#
is.integer(x)
#
# It's not a double precision object
#
is.double(x)
#
# Make a double precision, logical, complex and character object
#
double(5)
logical(5)
complex(5)
character(5)
#
# Or, we could just make a numeric vector
#
numeric(5)
#
# What is the mode?
#
x = double(5)
mode(x)
x = logical(5)
mode(x)
x = complex(5)
mode(x)
x = character(5)
mode(x)
x = numeric(5)
mode(x)
```

#### C.1.4 Index Referencing

R has a flexible system of index references:

```
x = 2*(1:10)
> x
[1]  2  4  6  8 10 12 14 16 18 20
> x[c(3,6,9)]
```

```
[1] 6 12 18
> y = x[c(3,6,9)]
> length(y)
[1] 3
> y
[1] 6 12 18
```

This makes it straightforward to create subvectors.

### C.1.5 More Vector Operations

The function `sort()` returns a sorted vector, while `sort.list()` returns a vector of indices which generates a sorted list. The function `rev()` returns a vector in reverse order, while `rank()` returns the ranks of the elements of a vector. By default, ties are represented by average ranks, although other options are available (see `help(rank)`).

```
> sort(c(2,4,5,3,2))
[1] 2 2 3 4 5
> x = c(2,4,5,3,2)
> sort(x)
[1] 2 2 3 4 5
> ind = sort.list(x)
> x[ind]
[1] 2 2 3 4 5
> rev(x)
[1] 2 3 5 4 2
> rank(x)
[1] 1.5 4.0 5.0 3.0 1.5
```

One useful function is `unique()` which returns the unique values of a vector:

```
> unique(c(1,2,2,3))
[1] 1 2 3
> unique(c(6,7,6,5,5,3,2,2,3))
[1] 6 7 5 3 2
> unique(c("Bob","bob","Mike", "Bob", "Bob "))
[1] "Bob"  "bob"  "Mike" "Bob "
```

Set operations can be performed on vectors. The `%in%` operator can be used to test for the presence of an element in a vector. It returns a logical value:

```
> 3 %in% c(1,2,3,4)
[1] TRUE
> 3 %in% c(1,2,4)
[1] FALSE
> "WNT3" %in% c("CCR5", "SFRP", "WNT3")
[1] TRUE
> "WNT3" %in% c("Ccr5", "Sfrp", "Wnt3")
[1] FALSE
```

The first argument may be a vector. In this case the test is applied to each element of the first vector, the second vector remaining fixed for each test. The result is therefore a logical vector equal in length to the first vector:

```
> c(2,3) %in% c(1,2,3,4)
[1] TRUE TRUE
> c(2,3,3,5,4,5,3) %in% c(1,2,3,4)
[1] TRUE TRUE TRUE FALSE TRUE FALSE TRUE
```

Formal set algebra is defined in R, in particular, `union()` ( $A \cup B$ ), `intersect()` ( $A \cap B$ ), `setdiff()` ( $A \cap B^c$ ). Note that `setdiff()` is the *asymmetric* set difference, not the *symmetric* set difference ( $(A \cap B^c) \cup (B \cap A^c)$ ).

```
> x = c(1,2,2,3,3,4)
> y = c(3,4,4,4,5,6,7,8,9)
> union(x,y)
[1] 1 2 3 4 5 6 7 8 9
> intersect(x,y)
[1] 3 4
> setdiff(x,y)
[1] 1 2
> setdiff(y,x)
[1] 5 6 7 8 9
```

The function `setequal()` ( $A = B$ ) tests for equality of vectors *regarded as sets*. It evaluates to a logical value, which is `TRUE` if and only if every element in one vector may be found in the other vector. Note that `unique()` returns a *vector* of the unique values in a vector, and so is ordered, that is, it is not a set.

```
> setequal(c(1,2),c(1,2))
[1] TRUE
> setequal(c(1,2),c(2,1))
[1] TRUE
> setequal(c(1,2),c(2,1,2,1,2,2))
[1] TRUE
```

Finally, `is.element(x,y)` is identical to `x %in% y`. See discussion below on *binary operators*.

Set operations may be performed on general types:

```
> gene.list.1 = c('ALDH4', 'A1AP2B1', 'BBC3', 'BCL2', 'CDC42BPA',
+                 'CDC42', 'CDCA7', 'CENPA', 'CDC42', 'CDCA7', 'CENPA',
+                 'CMC2', 'DHX58', 'DEXH', 'DIAPH3', 'DTL')
> gene.list.2 = c('A1AP2B1', 'BBC3', 'CDC42', 'CDCA7', 'CENPA',
+                 'CMC2', 'COL4A2', 'DCK', 'DHX58', 'DEXH', 'DIAPH3',
+                 'DTL')
> #
> # Which genes are in both lists?
> #
> intersect(gene.list.1, gene.list.2)
[1] "A1AP2B1" "BBC3"     "CDC42"    "CDCA7"    "CENPA"    "CMC2"
      "DHX58"   "DEXH"     "DIAPH3"   "DTL"
```

### C.1.6 Pattern Matching

R has quite extensive pattern matching capabilities. For example, using `grep()` we can specify a character pattern, then determine indices of a character vector containing that pattern:

```
> gene.list.1 = c('ALDH4', 'A1AP2B1', 'BBC3', 'BCL2', 'CDC42BPA',
+                 'CDC42', 'CDCA7', 'CENPA', 'CDC42', 'CDCA7', 'CENPA',
+                 'CMC2', 'DHX58', 'DEXH', 'DIAPH3', 'DTL')
> ind = grep('CDC', gene.list.1)
> ind
[1] 5 6 7 9 10
> gene.list.1[ind]
[1] "CDC42BPA" "CDC42"     "CDCA7"    "CDC42"    "CDCA7"
```

The function `grepl()` performs the same function but returns a logical vector:

```
ind = grepl('CDC', gene.list.1)
> ind
[1] FALSE FALSE FALSE FALSE TRUE  TRUE  TRUE FALSE
      TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
> gene.list.1[ind]
[1] "CDC42BPA" "CDC42"     "CDCA7"    "CDC42"    "CDCA7"
```

See also function `regexp()` and `gregexpr()`. Pattern replacement is done with functions `sub()` and `gsub()`. Most of these function support Perl-style regular expressions, by setting option `perl=TRUE`.

The functions `substr()` and `strsplit()` can be used to extract or replace patterns within a single character string.

### C.1.7 Managing Objects

There are a number of functions useful for managing R objects. The function `ls()` can be used to list currently available objects. It can be used without argument to list all objects:

```
> ls()
[1] "f"      "junk"   "junka"  "junkb"  "x"      "x1"    "x2"    "y"
```

Conditional lists can be made (see `help(ls)`).

Objects can be removed using the `rm()` command.

```
> ls()
[1] "f"      "junk"   "junka"  "junkb"  "x"      "x1"    "x2"    "y"
> rm(y)
> ls()
[1] "f"      "junk"   "junka"  "junkb"  "x"      "x1"    "x2"
```

Be aware that `ls()` will return a list of all objects, which, combined with `rm()` can be used to remove all objects:

```
> rm(list = ls())
> ls()
character(0)
```

If the previous example is examined carefully, it will be noticed that no warning prompt was given before R removed all objects.

## C.2 Data Structures in R

Besides vectors, data can be stored in matrices, as well as objects referred to as *arrays*, *lists* and *data frames*.

### C.2.1 Matrices

A matrix is essentially a vector with multiple indices. A two dimensional matrix can be created with the function:

```
matrix(data = NA, nrow = 1, ncol = 1, byrow = FALSE,
       dimnames = NULL)
```

The number of rows and columns are given explicitly by the `nrow` and `ncol` parameters. If a single value is specified for the data, all matrix elements will equal that value (with `NA` as default). It is also possible to input as data a vector of length  $nrow \times ncol$  to be copied sequentially into the matrix either by row or by column, as specified by the `byrow` option. Several examples follow:

```
> matrix(data=c(1,2,3,4,5,6), nrow=2, ncol=3, byrow=T)
 [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
> matrix(data=c(1,2,3,4,5,6), nrow=2, ncol=3, byrow=F)
 [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
> matrix(data=c(1,2,3,4,5,6), nrow=2, ncol=3)
 [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
> matrix(data=99, nrow=2, ncol=3)
 [,1] [,2] [,3]
[1,]   99   99   99
[2,]   99   99   99
> matrix(nrow=2, ncol=3)
 [,1] [,2] [,3]
[1,]   NA   NA   NA
[2,]   NA   NA   NA
```

The function `dim()` is used to either set or retrieve the dimensions of a matrix (row then column). For example:

```
> x = matrix(data=99, nrow=2, ncol=3)
> dim(x)
[1] 2 3
```

Dimensions can also be set for a vector of data:

```
> x = 1:12
> x
[1]  1  2  3  4  5  6  7  8  9 10 11 12
> dim(x) = c(4,3)
> x
[,1] [,2] [,3]
[1,]    1    5    9
```

```
[2,]    2    6   10
[3,]    3    7   11
[4,]    4    8   12
```

At this point, we can introduce the matrix transpose function `t()`:

```
> x = 1:12
> x
[1] 1 2 3 4 5 6 7 8 9 10 11 12
> dim(x) = c(3,4)
> x
 [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
> y = t(x)
> y
 [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
[3,]    7    8    9
[4,]   10   11   12
```

This allows us to impose  $4 \times 3$  dimensions onto a vector using by row sequence.

A diagonal matrix can be constructed from a vector with the `diag()` function:

```
> diag(rep(1,10))
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    1    0    0    0    0    0    0    0    0    0
[2,]    0    1    0    0    0    0    0    0    0    0
[3,]    0    0    1    0    0    0    0    0    0    0
[4,]    0    0    0    1    0    0    0    0    0    0
[5,]    0    0    0    0    1    0    0    0    0    0
[6,]    0    0    0    0    0    1    0    0    0    0
[7,]    0    0    0    0    0    0    1    0    0    0
[8,]    0    0    0    0    0    0    0    1    0    0
[9,]    0    0    0    0    0    0    0    0    1    0
[10,]   0    0    0    0    0    0    0    0    0    1
```

When the input to `diag()` is a matrix, the diagonal elements are returned:

```
> matrix(c(1,2,3,4),2,2)
 [,1] [,2]
```

```
[1,]    1    3
[2,]    2    4
> diag(matrix(c(1,2,3,4),2,2))
[1] 1 4
```

Matrix multiplication is carried out by the operator `%*%`:

```
> A = matrix(1:12,3,4)
> B = diag(c(1,2,2))
> A
[,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
> B
[,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    2    0
[3,]    0    0    2
> C = B%*%A
> C
[,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    4   10   16   22
[3,]    6   12   18   24
```

A matrix may be constructed by appending rows or columns using the `rbind()` or `cbind()` function:

```
> A = rbind(c(1,0,0,0), c(1,1,0,0), c(1,1,1,0), c(1,1,1,1))
> A
[,1] [,2] [,3] [,4]
[1,]    1    0    0    0
[2,]    1    1    0    0
[3,]    1    1    1    0
[4,]    1    1    1    1
> B = cbind(c(1,0,0,0), c(1,1,0,0), c(1,1,1,0), c(1,1,1,1))
> B
[,1] [,2] [,3] [,4]
[1,]    1    1    1    1
[2,]    0    1    1    1
[3,]    0    0    1    1
[4,]    0    0    0    1
```

The linear system of equations

$$b = Ax$$

for unknown vector  $x$  can be solved by the function `solve()`:

```
> A = rbind(c(1,0,0,0), c(1,1,0,0), c(1,1,1,0), c(1,1,1,1))
> A
[,1] [,2] [,3] [,4]
[1,]    1    0    0    0
[2,]    1    1    0    0
[3,]    1    1    1    0
[4,]    1    1    1    1
> b = 1:4
> b
[1] 1 2 3 4
> x = solve(A,b)
> x
[1] 1 1 1 1
> A%*%x
[,1]
[1,]    1
[2,]    2
[3,]    3
[4,]    4
> dim(x)
NULL
```

Note that in the operator `A%*%x` of the previous example, a matrix dimension  $4 \times 1$  was assumed for `x`, even though it is not a matrix. In this case, the dimension value of `x` is interpreted as an empty vector.

If we want just the inverse of a matrix we include only that matrix in `solve()` function:

```
> A = rbind(c(1,0,0,0), c(1,1,0,0), c(1,1,1,0), c(1,1,1,1))
> A
[,1] [,2] [,3] [,4]
[1,]    1    0    0    0
[2,]    1    1    0    0
[3,]    1    1    1    0
[4,]    1    1    1    1
> solve(A)
[,1] [,2] [,3] [,4]
[1,]    1    0    0    0
```

```
[2,]   -1    1    0    0
[3,]    0   -1    1    0
[4,]    0    0   -1    1
> solve(A) %*% A
 [,1] [,2] [,3] [,4]
[1,]    1    0    0    0
[2,]    0    1    0    0
[3,]    0    0    1    0
[4,]    0    0    0    1
```

### C.2.2 More on Index Subsets

The indexing of matrices is flexible. Recall that subvectors can be created from vectors using vectors of indices within the square brackets:

```
> x = seq(1,100,7)
> x
[1]  1  8 15 22 29 36 43 50 57 64 71 78 85 92 99
> length(x)
[1] 15
> x[c(3,6:9)]
[1] 15 36 43 50 57
> x[]
[1]  1  8 15 22 29 36 43 50 57 64 71 78 85 92 99
> x[NULL]
numeric(0)
```

In the preceding example, a subvector consisting of the 3rd, 6th, 7th, 8th and 9th element of a vector of length 15 is displayed. Leaving the square brackets empty yields the entire vector, while using `NULL` as the index vector yields a zero length vector.

The same principle applies to matrices and matrix like objects, except that each dimension is subsetted. If the index set is empty all elements of that dimension are included. This allows row and column vectors to be extracted from a matrix.

```
> A = cbind(c(1,26,1,1), c(7,76,7,7), c(4,43,4,4), c(3,39,3,3))
> A
 [,1] [,2] [,3] [,4]
[1,]    1    7    4    3
[2,]   26   76   43   39
[3,]    1    7    4    3
[4,]    1    7    4    3
> A[,2:3]
```

```
[,1] [,2]
[1,]    7    4
[2,]   76   43
[3,]    7    4
[4,]    7    4
> A[1:3,]
     [,1] [,2] [,3] [,4]
[1,]    1    7    4    3
[2,]   26   76   43   39
[3,]    1    7    4    3
> A[1,]
[1] 1 7 4 3
> A[,4]
[1] 3 39 3 3
> A[2:3,3:4]
     [,1] [,2]
[1,]   43   39
[2,]    4    3
> A[2:3,3]
[1] 43 4
```

Note that a column or row vector (or subvector) is not a matrix. That is  $A[2:3,3]$  yields a vector of length 2, and not a  $2 \times 1$  matrix.

Negative indices can be used to *remove* elements from a vector or matrix:

```
> x = c(3:10)
> x
[1] 3 4 5 6 7 8 9 10
> x[-1]
[1] 4 5 6 7 8 9 10
> x[-(4:7)]
[1] 3 4 5 10
```

Also, logical vectors may be used to subset indices.

```
> x = 1:4
> z = c(T,T,F,T)
> x[z]
[1] 1 2 4
> x[c(FALSE,TRUE)]
[1] 2 4
> x[c(FALSE,TRUE,TRUE)]
[1] 2 3
> x = 1:50
```

```
> x[x%%7==0]
[1] 7 14 21 28 35 42 49
```

In the last command, all positive integers not greater than 50 which are divisible by 7 have been enumerated.

Normally, logical vectors used in the way are of the same length as the vector. If this is not the case, R will recycle the logical vector up to the same length.

The function `which()` can be used to identify the indices of TRUE elements of a logical vector:

```
> which(c(T,T,F,T))
[1] 1 2 4
```

This function is very useful when vector Boolean expressions are used. Repeated the previous example:

```
> x = seq(0, 50, 7)
> x
[1] 0 7 14 21 28 35 42 49
> ind = which(x%%3 == 0)
> ind
[1] 1 4 7
> x[ind]
[1] 0 21 42
```

### C.2.3 Lists

A list is a labeled or ordered collection of any type of object. It has a number of uses, including the organization of function input and output, or the storage of irregular forms of data.

```
> input.obj = list(init.values = c(0.2, 0.1, 10), tolerance = 0.001,
maxIter = 100, memo.label = "July 3, 2013")
> input.obj
$init.values
[1] 0.2 0.1 10.0

$tolerance
[1] 0.001

$maxIter
[1] 100
```

```
$memo.label  
[1] "July 3, 2013"  
  
> input.obj[[1]]  
[1] 0.2 0.1 10.0  
> input.obj[[2]]  
[1] 0.001  
> input.obj[[3]]  
[1] 100  
> input.obj[[4]]  
[1] "July 3, 2013"  
>
```

Elements of a list can be addressed either by their label or by index (using double square brackets).

Lists of a specified length may be created in the following way

```
> xlist = vector('list',4)  
> xlist  
[[1]]  
NULL  
  
[[2]]  
NULL  
  
[[3]]  
NULL  
  
[[4]]  
NULL  
  
> xlist[[3]] = "3rd entry in list"  
> names(xlist) = paste('memo',1:4)  
> xlist  
$'memo 1'  
NULL  
  
$'memo 2'  
NULL  
  
$'memo 3'  
[1] "3rd entry in list"
```

```
$`memo 4'
NULL
```

As is the case for vectors, R offers considerable flexibility in constructing lists. The function `list()` with no argument creates a list of length 0. The length can be extended to k by making an assignment to the kth element.

```
> x = list()
> x
list()
> length(x)
[1] 0
> xlist = list()
> xlist
list()
> length(xlist)
[1] 0
> xlist[[4]] = '4th element'
> xlist
[[1]]
NULL

[[2]]
NULL

[[3]]
NULL

[[4]]
[1] "4th element"

>
```

The function `split()` creates a list by separating elements of a vector `x` by groups defined in `y`:

```
> x = c(1,2,3,4,5,6,7,8)
> y = rep(1:2,4)
> x
[1] 1 2 3 4 5 6 7 8
> y
[1] 1 2 1 2 1 2 1 2
> xy = split(x,y)
> xy
```

```
$ '1'
[1] 1 3 5 7

$ '2'
[1] 2 4 6 8
```

### C.2.4 Data Frames

The data frame is an important type of object in R. In statistical applications data often consists of samples of *records*, or collections of various forms of data in a fixed structure. This has a tabular form, but it need not be a matrix. In R is it taken to be a list of vectors of common length, in which the vectors may be of various types. These vectors can be assembled using the `data.frame()` command:

```
> patient.id = c(101,102,103)
> gender = c('M', 'M', 'F')
> bmi = c(103, NA, 131)
> bmi.data = data.frame(patient.id, gender, bmi)
> bmi.data
  patient.id gender bmi
1          101      M 103
2          102      M   NA
3          103      F 131
> bmi.data
  patient.id gender bmi
1          101      M 103
2          102      M   NA
3          103      F 131
> bmi.data[[1]]
[1] 101 102 103
> bmi.data}gender
[1] M M F
Levels: F M
>
```

Note that as a list the columns of a data frame can be accessed by label or by index.

### C.2.5 Factors

Note that in the previous example of Section C.2.4, although the vector `gender` was created as vector of character strings, within the data frame it is interpreted as a vector of `factors`, which is nonnumerical data assuming

one of a finite number of well defined `levels`. This is the standard way of representing *categorical data* in R (see Section 10.1), so it is important to understand the various conventions.

A character vector can (sometimes) be interpreted as a factor but, for example, an integer vector must be coerced.

```
> x = c(2,3,2,2,1,3,2)
> x = as.factor(x)
> x
[1] 2 3 2 2 1 3 2
Levels: 1 2 3
> attributes(x)
$levels
[1] "1" "2" "3"

$class
[1] "factor"
```

A vector of factors includes, as an object, the `levels` giving the possible values (categories). This means a level need not be represented in the vector:

```
> y = x[1:4]
> y
[1] 2 3 2 2
Levels: 1 2 3
```

Often, when a factor is an appropriate input, a character vector, or even a numerical vector can be used instead. However, to avoid ambiguity, it is best to create factors when appropriate.

### C.2.6 Arrays

An array object is a multidimensional array of elements of common type. A matrix is a two-dimensional array. However, a vector is not a one-dimensional array, because it has no dimension structure. It can be created with the `array()` function, in a manner similar to, but more general than, the `matrix()` function:

```
> array(data = 1:12, dim = c(2,6))
      [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    1    3    5    7    9   11
[2,]    2    4    6    8   10   12
> array(data = 1:12, dim = c(12))
[1] 1 2 3 4 5 6 7 8 9 10 11 12
```

```
> array(data = 1:12, dim = c(2,2,3))
, , 1

[,1] [,2]
[1,]    1    3
[2,]    2    4

, , 2

[,1] [,2]
[1,]    5    7
[2,]    6    8

, , 3

[,1] [,2]
[1,]    9   11
[2,]   10   12

> array(data = 1:12, dim = c(2,2,3))[1,2,1]
[1] 3
> array(data = 1:12, dim = c(2,2,3))[, , 2]
[,1] [,2]
[1,]    5    7
[2,]    6    8
```

Arrays with 3 or more dimensions can be created, and managed using the same type of indexing used by matrices, but with 3 or more indices.

### C.3 Labels for Data Structures

One very important feature of R is the assignment of labels to objects (vectors, matrices, lists, data frames). We will look at functions `names()`, `rownames()`, `colnames()`, `row.names()`, `dimnames()`. Note that `names()` is a generic function, meaning that its exact function depends on the type of object to which it is applied.

This is a good point at which to introduce the `paste()` function, which concatenates character strings

```
> paste("Elvis","Presley")
[1] "Elvis Presley"
> paste("Elvis","Presley",sep="-")
[1] "Elvis-Presley"
```

```
> paste("Elvis","Presley",sep="")
[1] "ElvisPresley"
```

The `sep` option defines the separator. The default is a single blank.  
Numbers are coerced to character strings, and vectors remain vectors:

```
> paste(1:6)
[1] "1" "2" "3" "4" "5" "6"
> paste(1:6)[2]
[1] "2"
> paste(1:6)[2:5]
[1] "2" "3" "4" "5"
```

Vector concatenation can be useful:

```
> paste('gene', 1:12, sep='')
[1] "gene1"   "gene2"   "gene3"   "gene4"   "gene5"   "gene6"   "gene7"
"gene8"   "gene9"   "gene10"  "gene11"  "gene12"
> paste('gene', 1:12, sep='-')
[1] "gene-1"  "gene-2"  "gene-3"  "gene-4"  "gene-5"  "gene-6"
"gene-7"  "gene-8"  "gene-9"  "gene-10" "gene-11"
[12] "gene-12"
```

### C.3.1 Vector Labels

The `names()` function can be used to set and access labels assigned to the elements of a vector. For example, a function might return an estimate, a standard error and a p-value for an associated hypothesis test. These three values can be placed in a single vector, and labels assigned to identify the values:

```
> x = c(20.3, 1.02, 0.023)
> names(x) = c("Est", "SE", "P-value")
> x
      Est       SE P-value
20.300 1.020 0.023
> names(x)
[1] "Est"       "SE"        "P-value"
> names(x)[2]
[1] "SE"
> names(x)[2] = "S.E."
> x
      Est     S.E. P-value
20.300 1.020 0.023
```

### C.3.2 Matrix and Array Labels

Matrix labels can be created, accessed changed in the much the same way using the `rownames()` and `colnames()` function:

```
> m = matrix(1:24,4,6)
> m
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    1    5    9   13   17   21
[2,]    2    6   10   14   18   22
[3,]    3    7   11   15   19   23
[4,]    4    8   12   16   20   24
> rownames(m) = paste('gene',1:4,sep=' ')
> colnames(m) = paste('subject',1:6,sep=' ')
> m
      subject1 subject2 subject3 subject4 subject5 subject6
gene1        1        5        9       13       17       21
gene2        2        6       10       14       18       22
gene3        3        7       11       15       19       23
gene4        4        8       12       16       20       24
> rownames(m) = paste('protein',1:4,sep=' ')
> m
      subject1 subject2 subject3 subject4 subject5 subject6
protein1      1        5        9       13       17       21
protein2      2        6       10       14       18       22
protein3      3        7       11       15       19       23
protein4      4        8       12       16       20       24
```

The combined row and column labels also exist as a single object, namely, a list of two vectors (consisting of the row and column labels), which can be managed with the `dimnames()` function:

```
> dimnames(m)
[[1]]
[1] "protein1" "protein2" "protein3" "protein4"

[[2]]
[1] "subject1" "subject2" "subject3" "subject4" "subject5" "subject6"

> dimnames(m)[[2]]
[1] "subject1" "subject2" "subject3" "subject4" "subject5" "subject6"
> dimnames(m)[[2]] = paste('patient',1:6,sep=' ')
> m
      patient1 patient2 patient3 patient4 patient5 patient6
```

protein1	1	5	9	13	17	21
protein2	2	6	10	14	18	22
protein3	3	7	11	15	19	23
protein4	4	8	12	16	20	24

For arrays of general dimension the `dimnames()` function can be used in much the same way it is used for matrices

```
> m = array(1:8, c(2,2,2))
> m
, , 1

[,1] [,2]
[1,]    1    3
[2,]    2    4

, , 2

[,1] [,2]
[1,]    5    7
[2,]    6    8

> dimnames(m) = list(c('black','white'),c('up','down'),c('left','right'))
+
> dimnames(m) = list(c('black','white'),c('up','down'),c('left','right'))
> m
, , left

      up down
black  1    3
white  2    4

, , right

      up down
black  5    7
white  6    8

> m[1,2,]
left right
      3    7
> m[1,,]
```

	left	right
up	1	5
down	3	7

### C.3.3 Labels for Lists and Data Frames

Labels for lists can be managed with the `name()` function:

```
> input.obj = list(init.values = c(0.2, 0.1, 10), tolerance = 0.001,
                   maxIter = 100, memo.label = "July 3, 2013")
> names(input.obj)
[1] "init.values" "tolerance"    "maxIter"      "memo.label"
> names(input.obj)[2]
[1] "tolerance"
> names(input.obj)[2] = "maxTolerance"
> input.obj
$init.values
[1] 0.2 0.1 10.0

$maxTolerance
[1] 0.001

$maxIter
[1] 100

$memo.label
[1] "July 3, 2013"
```

Note that the name originally used for the second element of the preceding list has been changed from "tolerance" to "maxTolerance".

A data frame is a ‘matrix-like object’, and so the functions `rownames()`, `colnames()` and `dimnames()` can be used:

```
> patient.id = c(101,102,103)
> gender = c('M', 'M', 'F')
> bmi = c(103, NA, 131)
> bmi.data = data.frame(patient.id, gender, bmi)
>
> rownames(bmi.data)
[1] "1" "2" "3"
> colnames(bmi.data)
[1] "patient.id" "gender"      "bmi"
> dimnames(bmi.data)
[[1]]
```

```
[1] "1" "2" "3"

[[2]]
[1] "patient.id" "gender"      "bmi"

> rownames(bmi.data) = paste('subject', 1:3)
> bmi.data
  patient.id gender bmi
subject 1      101     M 103
subject 2      102     M  NA
subject 3      103     F 131
>
```

A data frame is also a list, so `names()` may be used. It has the same effect as `colnames()`. The function `row.names()` is intended for data frames, but has for most purposes the same functionality as `rownames()`.

```
> names(bmi.data)
[1] "patient.id" "gender"      "bmi"
> row.names(bmi.data)
[1] "subject 1" "subject 2" "subject 3"
```

## C.4 Programming and Functions

R contains high-level programming capabilities. Commands entered on the command line can be included as lines in a program. The program can be invoked from an interactive editor (this is OS dependent) or run from a file using the `source()` function.

### C.4.1 Program Control

Program control relies on the reserved words `for`, `if`, `else`, `while`, `repeat`, `break`, `next`.

The syntax of `for` loops is somewhat different for R, consisting of a sequential assignment of elements of a vector to a variable. Formally it is given by

$$\text{for } (\text{var} \text{ in } \text{vector}) \text{expr}$$

for iterative evaluation of a single command, or

$$\text{for } (\text{var} \text{ in } \text{vector}) \{\text{code block}\}$$

for iterative evaluation of a block of code. Reserved words `break` or `next` can be used to terminate the loop, or to proceed immediately to the next iteration, respectively. A single expression example follows:

```
> x = c('A','B','C')
> y = rep(NA,3)
> for (i in 1:3) y[i] = paste(x[i],i,sep='')
> y
[1] "A1" "B2" "C3"
```

The iterated vector can be any type:

```
> x = c('A','B','C')
> y = NULL
> for (i in x) y = c(y,paste('[',i,']',sep=''))
> y
[1] "[A]" "[B]" "[C]"
```

The following example iterates a block of code enclosed in braces {}:

```
> x = 0
> for (i in 1:10) {
+   y = i^2+1
+   x = x + y
+ }
> x
[1] 395
> sum( (1:10)^2 + 1)
[1] 395
```

The loop is intended to evaluate  $\sum_{i=1}^{10} (i^2 + 1)$ , and succeeds in doing so.

The **if** reserved word allows condition executions of expressions or code blocks. The parenthesis contains a conditional expression. If true, the expression or block is evaluated.

```
> x = 3
> a1 = 0
> a2 = 0
> if (x == 3) {
+   a1 = 1
+   a2 = 1
+ }
> a1
[1] 1
> a2
[1] 1
```

The **else** reserved word specifies an expression or code block to evaluate if the condition is false

```
> x = 3
> a1 = 0
> a2 = 0
> if (x > 3) {
+   a1 = 1
+   a2 = 1
+ } else
+ {
+   a1 = 99
+   a2 = 99
+ }
> a1
[1] 99
> a2
[1] 99
```

### C.4.2 User Defined Functions

User defined functions are easy to make in R. The expression `function(...){...}` is assigned to an object, which becomes a *function object*. The parentheses includes all arguments, and the brackets include the block of code. The `return()` function is used to define the output:

```
> mean.and.variance = function(x) {
+
+   n = length(x)
+   meanx = sum(x)/n
+   varx = (sum(x^2) - (sum(x)^2)/n)/(n-1)
+   ans = c(meanx, varx)
+   names(ans) = c("mean", "variance")
+   return(ans)
+ }
>
> x = c(3,4,2,3,4,5,1,5)
> mean.and.variance(x)
  mean variance
3.375000 1.982143
> y = mean.and.variance(x)
> y
  mean variance
3.375000 1.982143
>
> mean(x)
```

```
[1] 3.375
> var(x)
[1] 1.982143
```

Needless to say, R already has functions which calculate means and variances.

One of the previous examples can be redesigned as a function:

```
> strange.sum = function(n) {
+ x = 0
+ for (i in 1:n) {
+   y = i^2+1
+   x = x + y
+ }
+ return(x)
+ }
> strange.sum(10)
[1] 395
```

Functions consisting of only a single expression do not require `return()` commands or braces (although braces can be used)

```
> mySum = function(x,y) {x + 2*y}
> mySum(2,4)
[1] 10
> mySum = function(x,y) x + 2*y
> mySum(2,4)
[1] 10
```

There is, of course, much more to this topic. Consult the R manual, or use `help('function')` (`function` is considered a reserved word).

### C.4.3 Functions and Environments

An *environment* is a *frame*, or collection of named R objects, and a pointer to an *enclosing environment*. It defines which R objects can be recognized for referencing. If an object is referenced, the current environment is searched. If it is not found, the enclosing environment is searched, and so on. The *global environment* `.GlobalEnv` (the user's workspace) is the first item on the search path, and is the current environment when R is started.

When a function is invoked, a new environment is created, and the calling environment becomes the enclosing environment. The function `environment()` displays the current environment.

```
> f = function(x) {environment()}
> environment()
<environment: R_GlobalEnv>
> f(0)
<environment: 0x109871e30>
```

An object created within a function is not the same object as an object in the enclosing environment.

```
> y = 5
> f = function(x) {
+   y = 99
+   return(y)
+ }
> f(0)
[1] 99
```

However, if an object is referenced but not found within an environment created by a function, it will search the enclosing environment.

```
> y = 5
> f = function(x) {
+   return(y)
+ }
> f(0)
[1] 5
> y = 6
> f(0)
[1] 6
```

Being aware of the *scope* of an object (in which environment an object is recognized) can be important when using a repository such as **bioconductor.org**. See <http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-scope.pdf> for more detail.

#### C.4.4 User Defined Binary Operators

A function can be expressed as a *binary operator* using the assignment

```
> "%myBinaryOperator%" = function(x, y) { ... }
> x %myBinaryOperator% y
```

Binary operations can then be placed directly into algebraic expressions:

```
> '%max%' = function(x,y) {max(x,y)}
> 5 %max% 4
[1] 5
```

## C.5 Vectorized Calculations

It is generally recommended in R that loops be avoided, and vectorized calculations used instead. For example, to evaluate  $\sum_{i=1}^{10}(x^2 + 1)$  we used a loop, but we also used a type of vectorized calculation to achieve the same effect:

```
> sum((1:10)^2 + 1)
[1] 395
```

Suppose we are given a numeric vector  $x$  and a second vector  $y$  of the same length which identifies groups:

```
> x = c(1,2,3,2,3,4)
> y = c(1,1,1,1,2,2)
```

We wish to calculate a mean for each group defined by  $y$ . We could write a general function to do this:

```
> group.means = function(x,y) {
+ 
+   group.names = unique(y)
+   n.group = length(group.names)
+   gmeans = rep(NA, n.group)
+   for (i in 1:n.group) gmeans[i] = mean(x[y==group.names[i]])
+   ans = list(group.names, gmeans)
+   names(ans) = c('group.names', 'group.means')
+   return(ans)
+ 
+ }
>
> group.means(x,y)
$group.names
[1] 1 2

$group.means
[1] 2.0 3.5
```

A much simpler method is to apply the function `tapply()`:

```
> tapply(x,y,mean)
  1   2
2.0 3.5
```

In this function, the first and second arguments are the data and the group labels (these can be of any type). The third argument is any function which

can be applied to a vector of the same type as `x`. There are several types of similar vectorized calculation functions. For example, `apply()` will apply a specified function to either the rows or columns of a matrix. See also `lapply()`, `sapply()`, `mapply()` and so on.

## C.6 File Input and Output

There are a number of methods for reading and writing data. R objects can be conveniently saved and restored using the `save()` and `load()` commands (see `help`). Tabular data is typically input using the `read.table()` command. A command such as

```
> new.data = read.table("myData.txt")
```

is usually intended to store a data frame in the variable `new.data` using data from the file `myData.txt`. There are many options, so documentation should be consulted. A data frame is save in a file using the `write.table()` command.

## C.7 Packages

All R functions are part of `packages`. The R distribution comes with the `base` package of standard R functions. All other packages are add-on packages and become a permanent part of the local installation (a number of add-on packages are already installed in the R distribution). An installed add-on package must be loaded into the R environment to be used during a session (even if it has already been installed). It is possible to define default packages, which can be listed with the following command:

```
> options("defaultPackages")
$defaultPackages
[1] "datasets"    "utils"        "grDevices"   "graphics"
      "stats"       "methods"
```

This can be changed, but to do so will require a more in depth understanding of the R startup procedure. See *An Introduction to R*.

To see packages which are installed use the command:

```
> library()
```

to obtain a display such as

```
Packages in library '/Library/Frameworks/R.framework/Versions
/3.0/Resources/library':
```

AnnotationDbi	Annotation Database Interface
base	The R Base Package
Biobase	Biobase: Base functions for Bioconductor
BiocGenerics	Generic functions for Bioconductor
BiocInstaller	Install/Update Bioconductor and CRAN Packages
bnlearn	Bayesian network structure learning, parameter learning and inference
boot	Bootstrap Functions (originally by Angelo Canty for S)
class	Functions for Classification
cluster	Cluster Analysis Extended Rousseeuw et al.
.	
.	
.	

To see packages which are loaded use the command

```
> search()
[1] ".GlobalEnv"      "package:linprog"   "package:lpSolve"
     "package:Matrix"    "tools:RGUI"       "package:stats"
     "package:graphics"
[8] "package:grDevices" "package:utils"     "package:datasets"
     "package:methods"   "Autoloads"       "package:base"
```

A default package is loaded automatically during the R startup. Otherwise a package may be loaded from the command line using the `library()` function. Sometimes a function will make use of another function belonging to a package which is generally not a default package. In this case a `require()` function may be included which will load the required function if it is installed, or otherwise issue a warning. Because R relies heavily on contributed packages, this feature is commonly encountered.

It is worth looking at the `help()` command in some more detail. A string argument not enclosed with quotes is assumed to be the name of an object (usually a function), while a string enclosed in quotes is a reserved word, such as `if`, `for` or `TRUE`. For example, the following commands:

```
> help(solve)
> help("for")
```

will provide detailed documentation on the function `solve()` and on the reserved word `for`, used in constructing `for` loops. If we want documentation on a specific package, we can use the option `package`:

```
> help(package = stats)
```

A list of packages available for installation is given using the command

```
> install.packages()
```

If the package to be installed is known, it becomes the argument to the function, enclosed in quotes:

```
> install.packages("bnlearn")
```

There are other ways of finding and installing packages, so see the documentation for more detail.

To take an example, suppose we are interested in linear programming. This is not a standard R function, but is available in an R repository (as are many numerical algorithms in fields other than statistics). If we don't know where to find an appropriate library we can use a command suitable for vague searches of documentation:

```
> help.search("linear programming")
```

This will yield documentation on packages including "linear programming" in specific features of the documentation (this can be specified by the user; see `help(help.search)` for more detail). In this case the packages `boot`, `linprog` and `lpSolve` are listed. We may then install, say, `linprog` using the command:

```
> install.packages("linprog")
trying URL 'http://lib.stat.cmu.edu/R/CRAN/bin/macosx/contrib
/3.0/linprog_0.9-2.tgz'
Content type 'application/x-gzip' length 33655 bytes (32 Kb)
opened URL
=====
downloaded 32 Kb
```

```
The downloaded binary packages are in
/var/folders/vs/v6dm307j09jfynfmgrpfplr0000gn/T
//RtmpWJwRiz/downloaded_packages
>
```

We can obtain documentation for the package, then load it, using command:

```
> library(help = linprog)
> library(linprog)
```

Because R has extensive repositories of contributed packages there will often be several alternatives, so it is usually advisable to do some comparisons.

Note that '?' and '??' are shortcuts for `help()` and `help.search()`. For example:

```
> ?quantile
> ???'linear programming'
> ??eigenvalues
```

Packages are sometimes updated. The function `update.packages()` will compare the local library with the appropriate repositories to determine packages for which updates are available. By default, the user is asked if the library should be updated. Many options are available, so consult `help(update.packages)`:

```
> update.packages()
cluster :
  Version 1.14.4 installed in /Library/Frameworks/R.framework
  /Versions/3.0/Resources/library
  Version 1.15.1 available at http://lib.stat.cmu.edu/R/CRAN
Update (y/N/c)? n
lattice :
  Version 0.20-24 installed in /Library/Frameworks/R.framework
  /Versions/3.0/Resources/library
  Version 0.20-27 available at http://lib.stat.cmu.edu/R/CRAN
Update (y/N/c)? n
lme4 :
  Version 1.0-6 installed in /Library/Frameworks/R.framework
  /Versions/3.0/Resources/library
  Version 1.1-5 available at http://lib.stat.cmu.edu/R/CRAN
Update (y/N/c)? c
cancelled by user
>
```

## C.8 Objects and Classes in R

We've so far used the term *object* to informally describe different forms of data structure in R (vectors, lists and so on). We note that R is a type of *object oriented* programming language, meaning that R objects possess specific forms of data structure, and are associated with various *methods*, or procedures which act on some or all of the the objects' data.

### C.8.1 Object Modes

An object can be characterized by their *mode*. An object is of *atomic* structure if all data elements are of the same type. The mode (identified by function `mode()`) is then that type of data:

```
> mode(c(1,2,3))
[1] "numeric"
> mode(c(T,F,F))
[1] "logical"
> mode(c('a','b','c'))
[1] "character"
```

A list is of mode `list`:

```
> list('a',c(2,3,2),TRUE)
[[1]]
[1] "a"

[[2]]
[1] 2 3 2

[[3]]
[1] TRUE
```

```
> mode(list('a',c(2,3,2),TRUE))
[1] "list"
```

and a function is of mode `function`

```
> mode(sum)
[1] "function"
```

### C.8.2 Object Classes

It is important to understand that there are several generations of object types in R. What they have in common is that their properties are defined by *attributes*, which are represented by a list, which can be accessed by the `attributes()` function:

```
> m = array(1:12,12)
> dimnames(m) = list(paste('c',1:12,sep=''))
> m
   c1  c2  c3  c4  c5  c6  c7  c8  c9  c10  c11  c12
   1   2   3   4   5   6   7   8   9   10  11  12
> attributes(m)
$dim
[1] 12

$dimnames
$dimnames[[1]]
```

```
[1] "c1"  "c2"  "c3"  "c4"  "c5"  "c6"  "c7"  "c8"
    "c9"  "c10" "c11" "c12"
```

Of most concern here are the S3 and S4 objects. The notion of R objects most familiar today are the S3 objects introduced around 1988. An S3 object has a *class attribute*, consisting (at least) of a character string which identifies a *class*. In effect, an object is an *instance* of a class (in the sense that  $x = 2.55$  is an instance of a real number). The object of the preceding example is of class *array*. Objects need only one class, but it may have several, in which case the class attribute would be a vector of class identifiers. When multiple classes are present, the order they are listed in the vector is important. The object can be considered a type of the first class listed, but it *inherits* from the subsequent classes in the order given. The `class()` function identifies the class of an object:

```
> class(m)
[1] "array"
```

### C.8.3 Generic Functions

Identifying classes permits the use of *generic functions*. These are functions which accept objects of varying classes, and whose functionality depends on that class. General purpose functions such as `print()`, `plot()` or `summary()` are typical generic functions. This is quite important in statistical modeling. For example, a fitted model may be a linear model, a logistic model or a Cox proportional hazards model, but each of these types share a common output and summary structure (ANOVA table, residual plot, and so on). In addition, model selection procedures (stepwise regression, cross validation) may be applied in much the same way to each. The functions which fit these models (in this case `lm()`, `glm()`, `coxph()`) typically output an object of a specific class (in this case `lm`, `glm`, `coxph`) but these may be input to one of several generic functions. Consider a simple example of a linear linear model, which regresses vector  $y$  onto vector  $x$ :

```
> x = c(1:10)
> y = c(2,4,3,4,5,6,1,2,3,5)
> fit = lm(y ~ x)
> class(fit)
[1] "lm"
> mode(fit)
[1] "list"
> fit
```

Call:

```
lm(formula = y ~ x)

Coefficients:
(Intercept)          x
3.26667      0.04242
```

the function `lm()` writes an `lm` class object into `fit`. To examine `fit` we can simply enter it into the command line, as though it were a variable. We get the formula used in the fit, as well as the regression coefficients. However, an `lm` object contains much more information than this. We can input `fit` into the generic function `summary()`:

```
> summary(fit)

Call:
lm(formula = y ~ x)

Residuals:
    Min     1Q Median     3Q    Max 
-2.56364 -1.14394  0.08485  1.14394  2.47879 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.26667   1.14186   2.861   0.0211 *  
x           0.04242   0.18403   0.231   0.8235    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.672 on 8 degrees of freedom
Multiple R-squared:  0.006599, Adjusted R-squared:  -0.1176 
F-statistic: 0.05315 on 1 and 8 DF,  p-value: 0.8235
```

We now get a much more complete summary of the model fit. In fact, we can get more information. The object attributes can be displayed using either the `names()` or the `attributes()` function (the `attributes()` function gives the class):

```
> attributes(fit)
$names
[1] "coefficients"  "residuals"       "effects"        "rank"        
[5] "fitted.values" "assign"         "df.residual"    "xlevels"      
[9] "call"           "terms"          "model"
```

```
$class
[1] "lm"

> names(fit)
[1] "coefficients"   "residuals"      "effects"       "rank"
[5] "fitted.values"  "assign"        "df.residual"   "xlevels"
[7] "qr"             "df.residual"   "xlevels"       "call"
[9] "terms"          "model"
```

For example, the model residuals can be accessed using the `residuals` attribute

```
fit$residuals
 1         2         3         4         5
 -1.3090909  0.6484848 -0.3939394  0.5636364  1.5212121
 6         7         8         9         10
 2.4787879 -2.5636364 -1.6060606 -0.6484848  1.3090909
>
```

For more description of the attributes use `help(lm)`

There exists a non-generic function `summary.lm()` which produces the same summary for an `lm` class. This is the standard nomenclature. If `myFunction()` is a generic function, then `myFunction.myClass()` is equivalent to `myFunction()` applied to an object of class `myClass`. Suppose an object `obj` has multiple classes `class1`, `class2`, `class3`, listed in that order. If `obj` in input to a generic function `fun()`, R will first search for function `fun.class1()`, then, if this is not found, `fun.class2()`, and so on. A generic function may have a default `fun.default()` which is used if no function for any of the objects classes is found.

```
> #
> # simulate a logistic regression model
> #
> x = c(1:10)
> y = rbinom(10, size = 1, prob = 0.5)
> fit.glm = glm(y ~ x, family = 'binomial')
> class(fit.glm)
[1] "glm" "lm"
```

The function `methods()` can be used either to display all generic functions available to a class, or all classes which may be input to a generic function:

```
> methods(summary)
[1] summary.aareg*           summary.aov*           summary.aovlist*
[4] summary.aspell*          summary.cch*          summary.connection*
```

```

[7] summary.coxph*           summary.coxph.penal*      summary.data.frame
[10] summary.Date             summary.default          summary.ecdf*
[13] summary.factor            summary.glm            summary.infl
[16] summary.lm                summary.loess*          summary.loglm*
[19] summary.manova            summary.matrix          summary.mlm
[22] summary.negbin*           summary.nls*           summary.packageStatus*
[25] summary.PDF_Dictionary*  summary.PDF_Stream*    summary.polr*
[28] summary.POSIXct            summary.POSIXlt        summary.ppr*
[31] summary.prcomp*           summary.princomp*     summary.proc_time
[34] summary.pyears*            summary.ratetable*    summary.rlm*
[37] summary.srcfile            summary.srcref         summary.stepfun
[40] summary.stl*               summary.survexp*      summary.survfit*
[43] summary.survfitms*        summary.survreg*       summary.table
[46] summary.tukeysmooth*      summary.XMLInternalDocument*


Non-visible functions are asterisked
> methods(class='lm')
[1] add1.lm*      alias.lm*      anova.lm      case.names.lm*  confint.lm*
[6] cooks.distance.lm* deviance.lm*  dfbeta.lm*    dfbetas.lm*   drop1.lm*
[11] dummy.coef.lm* effects.lm*    extractAIC.lm* family.lm*    formula.lm*
[16] hatvalues.lm   influence.lm*   kappa.lm     labels.lm*    logLik.lm*
[21] model.frame.lm  model.matrix.lm nobs.lm*     plot.lm       predict.lm
[26] print.lm       proj.lm*      qr.lm*       residuals.lm rstandard.lm
[31] rstudent.lm    simulate.lm*   summary.lm    variable.names.lm* vcov.lm*

```

```

Non-visible functions are asterisked
>
```

A non-visible function produces output, but this will not be displayed (it can be copied into an object).

#### C.8.4 User Defined Methods

A simple example of the construction of a method for a class based on an existing generic function follows:

```

> x = 2
> class(x) = "myClass"
> class(x)
[1] "myClass"
>
> # For some reason, we would like objects of this class
> # to be displayed 7 times.
>
> print.myClass = function(x) {rep(x,7)}
> print(x)
[1] 2 2 2 2 2 2 2
> y = unclass(x)
> print(y)
```

```
[1] 2
> methods(class='myClass')
[1] print.myClass
```

New generic functions can be constructed using the `UseMethod()` function.

### C.8.5 S4 (Formal) Classes

One problem with S3 classes is the rather informal nature of the identification of a class, which only requires setting a class attribute. In other words, classes are never really formally defined, which creates a problem when the same class of object is to be used by extensive collections of contributed packages. To address this problem S4 classes were introduced later in the 1990s. These require formal class definitions, which standardizes the attributes of an objects. This allows contributed applications to accept well defined standardized input. For this reason S4 classes are also referred to as *formal classes*.

The `setClass()` function is used to define a class (S3 classes have no corresponding functions):

```
> xyClass = setClass("xyClass", slots = c(x="numeric", y="numeric"))
>
> xy.obj1 = xyClass(x = c(1,2,3,4), y = c(8,7,6,5))
> xy.obj2 = new("xyClass", x = c(1,2,3,4), y = c(8,7,6,5))
>
> class(xy.obj1)
[1] "xyClass"
attr(,"package")
[1] ".GlobalEnv"
> class(xy.obj2)
[1] "xyClass"
attr(,"package")
[1] ".GlobalEnv"
```

The command `new()` can be used to create an instance of an S4 class. Also, because the identifier `xyClass` was used in the assignment `xyClass = setClass("xyClass", slots = c(x="numeric", y="numeric"))` a function `xyClass()` is created which can be used to create a class instance as shown above (the name need not be the same as the class).

The function `setMethod()` is used to create formal (S4) methods, however, the procedure for S3 classes may also be used:

```
> print.xyClass = function(obj) {rep(c(obj@x,obj@y),3)}
> print(xy.obj1)
```

See <http://cran.r-project.org/doc/contrib/Genolini-S4tutorialV0-5en.pdf> or <http://www.bioconductor.org/help/course-materials/2013/CSAMA2013/friday/afternoon/S4-tutorial.pdf> for more detail.

### C.8.6 Testing and Coercion of Object Types

It is important in an object oriented environment to be able to determine the type of an object, and to change it if necessary (this is generally referred to as coercion). These types of functions generally have names resembling `is.thing` or `as.thing`. For example:

```
> is.vector(c(2,3,3,4))
[1] TRUE
> is.matrix(c(2,3,3,4))
[1] FALSE
> as.matrix(c(2,3,3,4))
     [,1]
[1,]    2
[2,]    3
[3,]    3
[4,]    4
> is.na(NA)
[1] TRUE
> is.na(3)
[1] FALSE
> is.null(34)
[1] FALSE
> is.null(NULL)
[1] TRUE
> #
> # Is it an S4 class object?
> #
> xyClass = setClass("xyClass", slots = c(x="numeric", y="numeric"))
> xy.obj1 = xyClass(x = c(1,2,3,4), y = c(8,7,6,5))
> isS4(xy.obj1)
[1] TRUE
> isS4(c(2,3,3,4))
[1] FALSE
```

## Appendix D

# Methodological Summary

We make use of the following distributions:

$$\begin{aligned} Z &\sim N(0, 1) \\ T_\nu &\sim t\text{-distribution with } \nu \text{ degrees of freedom} \\ \chi_\nu^2 &\sim \chi^2\text{-distribution with } \nu \text{ degrees of freedom} \\ F_{\nu_1, \nu_2} &\sim F\text{-distribution with } \nu_1 \text{ and } \nu_2 \text{ numerator and denominator degrees of freedom} \end{aligned}$$

Critical values are denoted:

$$\begin{aligned} z_\alpha &\text{satisfies } P(Z > z_\alpha) = \alpha \\ t_{\nu, \alpha} &\text{satisfies } P(T_\nu > t_{\nu, \alpha}) = \alpha \\ \chi_{\nu, \alpha}^2 &\text{satisfies } P(\chi_\nu^2 > \chi_{\nu, \alpha}^2) = \alpha \\ F_{\nu_1, \nu_2, \alpha} &\text{satisfies } P(F_{\nu_1, \nu_2} > F_{\nu_1, \nu_2, \alpha}) = \alpha \end{aligned}$$

### D.1 Diagnostic Testing

Baye's theorem is given by the following argument:

$$\begin{aligned} P(A | E) &= \frac{P(E | A)P(A)}{P(E)} \\ &= \frac{P(E | A)P(A)}{P(E | A)P(A) + P(E | A^c)P(A^c)} \end{aligned}$$

The *odds* of an event is defined as

$$\begin{aligned} Odds(A) &= \frac{P(A)}{P(A^c)} \\ &= \frac{P(A)}{1 - P(A)}. \end{aligned}$$

$$\begin{aligned} Odds(A | E) &= \frac{P(A | E)}{P(A^c | E)} \\ &= \frac{P(A | E)}{1 - P(A | E)}. \end{aligned}$$

Then a form of Baye's theorem expressed in term of odds is given by

$$\begin{aligned} Odds(A | E) &= \frac{P(A | E)}{P(A^c | E)} \\ &= \frac{P(E | A)P(A)}{P(E)} \times \frac{P(E)}{P(E | A^c)P(A^c)} \\ &= \frac{P(E | A)}{P(E | A^c)} \times \frac{P(A)}{P(A^c)} \\ &= \frac{P(E | A)}{P(E | A^c)} \times Odds(A). \end{aligned}$$

We define the likelihood ratio to be

$$LR = \frac{P(E | A)}{P(E | A^c)}.$$

Baye's theorem, in this form, states

$$Odds(A | E) = LR \times Odds(A).$$

Suppose we are given outcomes

$$\begin{aligned} O_- &= \{ \text{the patient does not have the condition} \} \\ O_+ &= \{ \text{the patient has the condition} \} \\ T_- &= \{ \text{the patient tests negative} \} \\ T_+ &= \{ \text{the patient tests positive} \} \end{aligned}$$

Clearly,  $O_-^c = O_+$  and  $T_-^c = T_+$ , so that  $P(O_-) + P(O_+) = 1$  and  $P(T_-) + P(T_+) = 1$ .

All possibilities are given in the following table:

Table D.1: Outcomes of diagnostic testing

		Condition	
		Positive	Negative
Test	Positive	TP	FP
	Negative	FN	TN

where

$$\begin{aligned} \text{TP} &= T_+ \cap O_+ = \text{True Positive} \\ \text{FP} &= T_+ \cap O_- = \text{False Positive} \\ \text{TN} &= T_- \cap O_- = \text{True Negative} \\ \text{FN} &= T_- \cap O_+ = \text{False Negative}. \end{aligned}$$

$$\begin{aligned} \text{sensitivity} &= \frac{TP}{TP + FN} = P(T_+ | O_+) \\ \text{specificity} &= \frac{TN}{TN + FP} = P(T_- | O_-) \\ \text{positive predictive value (PPV)} &= \frac{TP}{TP + FP} == P(O_+ | T_+) \\ \text{negative predictive value (NPV)} &= \frac{TN}{TN + FN} == P(O_- | T_-). \end{aligned}$$

If we let  $N = TP + FP + TN + FN$  (the total number of entries in Table D.1 we can also calculate the marginal probabilities:

$$\begin{aligned} P(O_-) &= \frac{FP + TN}{N} \\ P(O_+) &= \frac{TP + FN}{N} \\ P(T_-) &= \frac{FN + TN}{N} \\ P(T_+) &= \frac{TP + FP}{N}. \end{aligned}$$

If prevalence is given (that is, not that obtained from a given contingency table), so that

$$prev = P(O^+),$$

then

$$\begin{aligned} PPV &= P(O_+ | T_+) \\ &= \frac{P(T_+ | O_+)P(O_+)}{P(T_+ | O_+)P(O_+) + P(T_+ | O_-)P(O_-)} \\ &= \frac{sens \times prev}{sens \times prev + (1 - spec) \times (1 - prev)} \end{aligned}$$

and

$$\begin{aligned} NPV &= P(O_- | T_-) \\ &= \frac{P(T_- | O_-)P(O_-)}{P(T_- | O_-)P(O_-) + P(T_- | O_+)P(O_+)} \\ &= \frac{spec \times (1 - prev)}{spec \times (1 - prev) + (1 - sens) \times prev} \end{aligned}$$

## D.2 Odds Ratios

We are given an outcome  $O_+$  and two conditions  $G_1, G_2$ . The odds ratio for  $O_+$  is defined as

$$OR = \frac{Odds(O_+ | G_1)}{Odds(O_+ | G_2)} = \frac{P(O_+ | G_1)/(1 - P(O_+ | G_1))}{P(O_+ | G_2)/(1 - P(O_+ | G_2))}.$$

Statistically, the  $OR$  can proceed by using a  $2 \times 2$  contingency table:

	$O_+$	$O_-$	
$G_1$	$n_{11}$	$n_{12}$	$R_1$
$G_2$	$n_{21}$	$n_{22}$	$R_2$
Total	$C_1$	$C_2$	$n$

The formula for the estimate of  $OR$  is simply

$$\hat{OR} = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

Inference proceeds by a natural log transformation  $\log(OR)$  of  $OR$ , which has standard error

$$SE(\log(OR)) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}.$$

An approximate  $(1 - \alpha)100\%$  confidence interval for  $\log(OR)$  is therefore

$$CI_{1-\alpha} = \log(\hat{OR}) \pm z_{\alpha/2} SE(\log(OR))$$

## D.3 Binomial Continuity Correction

If  $X_{bin} \sim bin(n, p)$  and  $X_{norm} \sim N(np, np(1 - p))$  then we have approximation

$$P(X_{bin} = k) \approx P(k - 0.5 \leq X_{norm} \leq k + 0.5).$$

This means the CDF of  $X_{bin}$  should use the approximation

$$F_{X_{bin}}(k) = P(X_{bin} \leq k) \approx P(X_{norm} \leq k + 0.5) = F_{X_{norm}}(k + 0.5) \quad (\text{D.1})$$

## D.4 Single Population Mean

We assume that there is a population of measurements with a true population mean  $\mu$ . We take a random sample from this population of size  $n$ , and accept the resulting sample mean  $\bar{X}_n$  as an estimate of  $\mu$ . Suppose

that the true population variance is  $\sigma^2$ . Then we may define a  $(1 - \alpha)100\%$  confidence interval to be

$$CI_{1-\alpha} = \bar{X}_n \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

If  $\sigma^2$  is unknown, then we substitute sample variance  $S$  for  $\sigma$  and  $t_{n-1,\alpha/2}$  for  $z_{\alpha/2}$  giving

$$CI_{1-\alpha} = \bar{X}_n \pm t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}.$$

Hypothesis tests use statistics:

$$Z_{obs} = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} (\sigma^2 \text{ known}), \quad T_{obs} = \frac{\bar{X}_n - \mu_0}{S/\sqrt{n}} (\sigma^2 \text{ unknown}).$$

1. One sided, lower tailed hypothesis

$$\begin{aligned} H_o &: \mu \geq \mu_0 \\ H_a &: \mu < \mu_0 \end{aligned}$$

If  $\sigma^2$  known: reject  $H_o$  if  $Z_{obs} \leq -z_\alpha$ , with P-value  $\alpha_{obs} = P(Z \leq Z_{obs})$ . If  $\sigma^2$  unknown: reject  $H_o$  if  $T_{obs} \leq -t_{n-1,\alpha}$ , with P-value  $\alpha_{obs} = P(T_{n-1} \leq T_{obs})$ .

2. One sided, upper tailed hypothesis

$$\begin{aligned} H_o &: \mu \leq \mu_0 \\ H_a &: \mu > \mu_0 \end{aligned}$$

If  $\sigma^2$  known: reject  $H_o$  if  $Z_{obs} \geq z_\alpha$ , with P-value  $\alpha_{obs} = P(Z \geq Z_{obs})$ . If  $\sigma^2$  unknown: reject  $H_o$  if  $T_{obs} \geq t_{n-1,\alpha}$ , with P-value  $\alpha_{obs} = P(T_{n-1} \geq T_{obs})$ .

3. Two sided hypothesis

$$\begin{aligned} H_o &: \mu = \mu_0 \\ H_a &: \mu \neq \mu_0 \end{aligned}$$

If  $\sigma^2$  known: reject  $H_o$  if  $|Z_{obs}| \geq z_{\alpha/2}$ , with P-value  $\alpha_{obs} = 2P(Z < -|Z_{obs}|)$ . If  $\sigma^2$  unknown: reject  $H_o$  if  $|T_{obs}| \geq t_{n-1,\alpha/2}$ , with P-value  $\alpha_{obs} = 2P(T_{n-1} < -|T_{obs}|)$ .

## D.5 Difference in Population Means

Independent samples have summaries:

	Pop'n 1	Pop'n 2
Population mean	$\mu_1$	$\mu_2$
Population variance	$\sigma_1^2$	$\sigma_2^2$
Sample size	$n_1$	$n_2$
Sample mean	$\bar{X}_1$	$\bar{X}_2$
Sample variance	$S_1^2$	$S_2^2$
Pooled Variance	$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$	

If the samples are paired,  $n = n_1 = n_2$ , and we calculate paired differences  $D_i = X_{2i} - X_{1i}$ ,  $i = 1, \dots, n$ . Then  $\bar{D}$  and  $S_D^2$  are the sample mean and sample variance of the differences  $D_1, \dots, D_n$ .

The degrees of freedom used in Welch's *t*-test is given by

$$\nu_W = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}}.$$

We consider 4 cases:

**Case 1** Samples are independent. Both population variances  $\sigma_1^2$  and  $\sigma_2^2$  are known.

**Case 2** Samples are independent. The population variances are unknown, but we assume they are equal, so that  $\sigma_1^2 = \sigma_2^2$ .

**Case 3** Samples are independent. The population variances are unknown and we cannot assume that they are equal.

**Case 4** Samples are paired.

1. One sided, lower tailed hypothesis

$$\begin{aligned} H_o &: \mu_2 \geq \mu_1 \\ H_a &: \mu_2 < \mu_1 \end{aligned}$$

We are looking for evidence that  $\mu_2$  is **less than**  $\mu_1$ .

2. One sided, upper tailed hypothesis

$$\begin{aligned} H_o &: \mu_2 \leq \mu_1 \\ H_a &: \mu_2 > \mu_1 \end{aligned}$$

We are looking for evidence that  $\mu_2$  is **greater than**  $\mu_1$ .

### 3. Two sided hypothesis

$$\begin{aligned} H_0 &: \mu_2 = \mu_1 \\ H_a &: \mu_2 \neq \mu_1 \end{aligned}$$

We are looking for evidence that  $\mu_2$  is **not equal to**  $\mu_1$ .

The following table summarizes, for each case, the procedures for constructing level  $1 - \alpha$  confidence intervals for  $\mu_2 - \mu_1$ , and for testing hypotheses (rejection region and P-value  $\alpha_{obs}$ ).

Case	$\nu$	$CI$	Test Statistics	Lower Tail	Upper Tail	Two sided
1	-	$\bar{X}_2 - \bar{X}_1$ $\pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$Z_{obs} = \frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	$Z_{obs} \leq -z_\alpha$ $P(Z \leq Z_{obs})$	$Z_{obs} \geq z_\alpha$ $P(Z \geq Z_{obs})$	$ Z_{obs}  \geq z_{\alpha/2}$ $2P(Z \leq - Z_{obs} )$
2	$n_1 + n_2 - 2$	$\bar{X}_2 - \bar{X}_1$ $\pm t_{\nu, \alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$	$T_{obs} = \frac{\bar{X}_2 - \bar{X}_1}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	$T_{obs} \leq -t_{\nu, \alpha}$ $P(T_\nu \leq T_{obs})$	$T_{obs} \geq t_{\nu, \alpha}$ $P(T_\nu \geq T_{obs})$	$ T_{obs}  \geq t_{\nu, \alpha/2}$ $2P(T_\nu \leq - T_{obs} )$
3	$\nu_W$	$\bar{X}_2 - \bar{X}_1$ $\pm t_{\nu, \alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$	$T_{obs} = \frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$	$T_{obs} \leq -t_{\nu, \alpha}$ $P(T_\nu \leq T_{obs})$	$T_{obs} \geq t_{\nu, \alpha}$ $P(T_\nu \geq T_{obs})$	$ T_{obs}  \geq t_{\nu, \alpha/2}$ $2P(T_\nu \leq - T_{obs} )$
4	$n - 1$	$\bar{D}$ $\pm t_{\nu, \alpha/2} \frac{S_D}{\sqrt{n}}$	$T_{obs} = \frac{\bar{D}}{S_D / \sqrt{n}}$	$T_{obs} \leq -t_{\nu, \alpha}$ $P(T_\nu \leq T_{obs})$	$T_{obs} \geq t_{\nu, \alpha}$ $P(T_\nu \geq T_{obs})$	$ T_{obs}  \geq t_{\nu, \alpha/2}$ $2P(T_\nu \leq - T_{obs} )$

## D.6 Inference for Population Proportions

Suppose the proportion in a population of a certain type is  $p$ . To estimate  $p$  we take a random sample of size  $n$  from the population. If  $\hat{p}$  is the proportion in the sample of the type of interest, then this serves as an estimate of  $p$ . In particular,  $\hat{p} = X/n$  where  $X \sim bin(n, p)$ . Furthermore, a consequence of the Central Limit Theorem is that, approximately

$$\hat{p} \sim N \left( p, \frac{p(1-p)}{n} \right).$$

This means that the standard deviation of  $\hat{p}$  is

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}.$$

Of course, if we are trying to estimate  $p$  that means we don't know its' value. We can, however, approximate the standard deviation by substituting  $\hat{p}$  for  $p$ , to get

$$\hat{\sigma}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

This leads to a level  $(1 - \alpha)100\%$  confidence interval for  $p$  given by

$$CI_{\alpha/2} = \hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}.$$

For hypothesis testing we again have three types of hypotheses. Suppose we set  $p_0$  to be some hypothetical population proportion, and let  $p$  be the true population proportion.

1. One sided, lower tailed hypothesis

$$\begin{aligned} H_o &: p \geq p_0 \\ H_a &: p < p_0 \end{aligned}$$

We are looking for evidence that  $p$  is **less than**  $p_0$ .

2. One sided, upper tailed hypothesis

$$\begin{aligned} H_o &: p \leq p_0 \\ H_a &: p > p_0 \end{aligned}$$

We are looking for evidence that  $p$  is **greater than**  $p_0$ .

3. Two sided hypothesis

$$\begin{aligned} H_o &: p = p_0 \\ H_a &: p \neq p_0 \end{aligned}$$

We are looking for evidence that  $p$  is **not equal to**  $p_0$ .

We then have the test statistic

$$Z_{obs} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

Note that we use the hypothetical value  $p_0$  in the test statistic.

There will be three different forms of the rejection region depending on the form of the hypothesis.

1. One sided, lower tailed hypothesis

Reject  $H_o$  if  $Z_{obs} \leq -z_\alpha$

2. One sided, upper tailed hypothesis

Reject  $H_o$  if  $Z_{obs} \geq z_\alpha$

3. Two sided hypothesis

$$\text{Reject } H_o \text{ if } |Z_{obs}| \geq z_{\alpha/2}$$

where  $z_\alpha$  is the  $\alpha$  critical value from a standard normal distribution.

The observed significance levels are calculated according to the following rules.

1. One sided, lower tailed hypothesis

$$\alpha_{obs} = P(Z \leq Z_{obs})$$

2. One sided, upper tailed hypothesis

$$\alpha_{obs} = P(Z \geq Z_{obs})$$

3. Two sided hypothesis

$$\alpha_{obs} = 2P(Z \leq -|Z_{obs}|)$$

where we assume  $Z \sim N(0, 1)$ .

For smaller sample sizes we can employ the continuity correction:

$$Z_{obs} = \frac{X + 0.5 - np_0}{\sqrt{np_0(1 - p_0)}},$$

for use in  $P(Z \leq Z_{obs})$ .

## D.7 Inference for Two Population Proportions

Assume that the population proportions of interest are  $p_1$  and  $p_2$ , and that random samples of size  $n_1$  and  $n_2$  are selected from each. Furthermore, suppose that the proportions of interest observed in the two samples are  $\hat{p}_1$  and  $\hat{p}_2$ . We further assume that the two samples were collected independently. Then a level  $(1 - \alpha)100\%$  confidence interval for  $p_2 - p_1$  is given by

$$CI_{1-\alpha} = \hat{p}_2 - \hat{p}_1 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}.$$

For hypothesis testing we again have three types of hypotheses. Suppose we set  $p_1$  and  $p_2$  to be the two population proportions. We then have the following hypotheses.

1. One sided, lower tailed hypothesis

$$\begin{aligned} H_o &: p_2 \geq p_1 \\ H_a &: p_2 < p_1 \end{aligned}$$

We are looking for evidence that  $p_2$  is **less than**  $p_1$ .

2. One sided, upper tailed hypothesis

$$\begin{aligned} H_o &: p_2 \leq p_1 \\ H_a &: p_2 > p_1 \end{aligned}$$

We are looking for evidence that  $p_2$  is **greater than**  $p_1$ .

3. Two sided hypothesis

$$\begin{aligned} H_o &: p_2 = p_1 \\ H_a &: p_2 \neq p_1 \end{aligned}$$

We are looking for evidence that  $p_2$  is **not equal to**  $p_1$ .

Note that for the null hypothesis we may set  $p_0 = p_1 = p_2$ . Since we usually construct the test statistic to have a certain distribution assuming that  $H_o$  is true, it makes sense to combine the two samples to form a single estimate of  $p_0$ , referred to as a **pooled** estimate of  $p_0$ . This is given by

$$\hat{p}_0 = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

We then have the test statistic

$$Z_{obs} = \frac{\hat{p}_2 - \hat{p}_1}{\sqrt{\hat{p}_0(1 - \hat{p}_0) \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Note that we use the pooled estimate  $\hat{p}_0$  in the test statistic.

There will be three different forms of the rejection region depending on the form of the hypothesis.

1. One sided, lower tailed hypothesis

Reject  $H_o$  if  $Z_{obs} \leq -z_\alpha$

2. One sided, upper tailed hypothesis

Reject  $H_o$  if  $Z_{obs} \geq z_\alpha$

3. Two sided hypothesis

$$\text{Reject } H_o \text{ if } |Z_{obs}| \geq z_{\alpha/2}$$

where  $z_\alpha$  is the  $\alpha$  critical value from a standard normal distribution.

Given this test statistic, the observed significance levels are calculated according to the following rules.

1. One sided, lower tailed hypothesis

$$\alpha_{obs} = P(Z \leq Z_{obs})$$

2. One sided, upper tailed hypothesis

$$\alpha_{obs} = P(Z \geq Z_{obs})$$

3. Two sided hypothesis

$$\alpha_{obs} = 2P(Z \leq -|Z_{obs}|)$$

where we assume  $Z \sim N(0, 1)$ .

## D.8 Sample Size Estimates (Population Mean)

Recall that the confidence interval for a population mean, given population variance  $\sigma^2$  is

$$CI_{1-\alpha} = \bar{X}_n \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

The margin of error is

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Now suppose before we collect the sample we decide that the margin of error should be  $E_o$ , and the confidence level should be  $(1 - \alpha)100\%$ . We can use the previous expression to determine what the sample size should be, giving

$$n = \left( z_{\alpha/2} \frac{\sigma}{E_o} \right)^2$$

as the required sample size. As a technical note, the  $n$  calculated will not generally be an integer. In this case we would always round up, as opposed to rounding to the nearest integer. This way, we ensure that the confidence level reported is not overestimated.

Of course, we would rarely know the actual value  $\sigma^2$ , so we would have to substitute an estimate  $\hat{\sigma}^2$ , giving

$$n \approx \left( z_{\alpha/2} \frac{\hat{\sigma}}{E_o} \right)^2.$$

If we had a previous study, or some other prior knowledge, we could rely on that for  $\hat{\sigma}^2$ . Failing that, a reasonable alternative is to do an initial **pilot study**. This would be a small sample whose primary purpose would be to obtain an estimate of  $\hat{\sigma}^2$ . An estimate of the required sample size for a fixed margin of error could then be obtained, and the sample then completed.

## D.9 Sample Size Estimates (Population Proportion)

The appropriate sample size required to estimate a population proportion can also be estimated using similar reasoning, except that there are some technical differences. Recall that the level  $(1 - \alpha)100\%$  confidence interval is given by

$$CI_{\alpha/2} = \hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

so that the margin of error is

$$E = z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}.$$

If this expression is rearranged we get

$$n = \hat{p}(1 - \hat{p}) \left( \frac{z_{\alpha/2}}{E} \right)^2.$$

A conservative estimate of  $n$  is obtained by replacing  $\hat{p}$  with  $1/2$ . If it is anticipated that  $\hat{p}$  will be much smaller than  $1/2$  substitute a plausible upper bound  $p^*$  for  $\hat{p}$ .

## D.10 Sample Size Estimates (Two Population Means)

Recall that the confidence interval for a difference in population means, given population variances  $\sigma_1^2$  and  $\sigma_2^2$ , and sample sizes  $n_1$  and  $n_2$  is

$$CI_{1-\alpha} = \bar{X}_2 - \bar{X}_1 \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

The margin of error is therefore

$$E = z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

We can, in principle, predict  $E$  for any configuration  $(\sigma_1^2, \sigma_2^2, n_1, n_2)$ . We may specify that  $n = n_1 = n_2$ , which is referred to as a *balanced design*. In this case we can obtain the formula

$$n \approx \left( \frac{z_{\alpha/2}}{E} \right)^2 (\sigma_1^2 + \sigma_2^2).$$

## D.11 Inference for Variances

If we are given a sample  $X_1, \dots, X_n$  from  $N(\mu, \sigma^2)$ , a level  $(1 - \alpha)$  confidence interval for  $\sigma^2$  is given by

$$\frac{S^2}{\left( \chi_{n-1, \alpha/2}^2 \right) / (n-1)} < \sigma^2 < \frac{S^2}{\left( \chi_{n-1, 1-\alpha/2}^2 \right) / (n-1)}.$$

and a level  $(1 - \alpha)$  upper confidence bound is given by

$$\sigma^2 < \frac{S^2}{\left( \chi_{n-1, 1-\alpha}^2 \right) / (n-1)}$$

where  $S^2$  is the sample variance.

Suppose we have two independent normally distributed populations:

	Pop'n 1	Pop'n 2
Population mean	$\mu_1$	$\mu_2$
Population variance	$\sigma_1^2$	$\sigma_2^2$
Sample size	$n_1$	$n_2$
Sample mean	$\bar{X}_1$	$\bar{X}_2$
Sample variance	$S_1^2$	$S_2^2$

Except that now we are concerned with hypotheses of the form

1. One sided, lower tailed hypothesis

$$\begin{aligned} H_o : \sigma_1^2 &\geq \sigma_2^2 \\ H_a : \sigma_1^2 &< \sigma_2^2 \end{aligned}$$

We are looking for evidence that  $\sigma_1^2$  is **less than**  $\sigma_2^2$ .

2. One sided, upper tailed hypothesis

$$\begin{aligned} H_o : \sigma_1^2 &\leq \sigma_2^2 \\ H_a : \sigma_1^2 &> \sigma_2^2 \end{aligned}$$

We are looking for evidence that  $\sigma_1^2$  is **greater than**  $\sigma_2^2$ .

3. Two sided hypothesis

$$\begin{aligned} H_o : \sigma_2^2 &= \sigma_1^2 \\ H_a : \sigma_2^2 &\neq \sigma_1^2 \end{aligned}$$

We are looking for evidence that  $\sigma_1^2$  is **not equal to**  $\sigma_2^2$ .

Use test statistic

$$F_{obs} = \frac{S_1^2}{S_2^2}.$$

There will be three different forms of the rejection region depending on the form of the hypothesis.

1. One sided, lower tailed hypothesis

Reject  $H_o$  if  $F_{obs} \leq F_{n_1-1, n_2-1, 1-\alpha}$

2. One sided, upper tailed hypothesis

Reject  $H_o$  if  $F_{obs} \geq F_{n_1-1, n_2-1, \alpha}$

3. Two sided hypothesis

Reject  $H_o$  if  $F_{obs} \leq F_{n_1-1, n_2-1, 1-\alpha/2}$  or  $F_{obs} \geq F_{n_1-1, n_2-1, \alpha/2}$ .

Note that lower tailed critical values can be obtained from the formula

$$F_{\nu_1, \nu_2, 1-\alpha} = \frac{1}{F_{\nu_2, \nu_1, \alpha}}.$$

Given this test statistic, the observed significance levels are calculated according to the following rules.

1. One sided, lower tailed hypothesis

$$\alpha_{obs} = P(F_{n_1-1, n_2-1} \leq F_{obs})$$

2. One sided, upper tailed hypothesis

$$\alpha_{obs} = P(F_{n_1-1, n_2-1} \geq F_{obs})$$

3. Two sided hypothesis

$$\alpha_{obs} = 2 \min(P(F_{n_1-1, n_2-1} \leq F_{obs}), P(F_{n_1-1, n_2-1} \geq F_{obs})).$$

## D.12 Goodness of Fit Tests

Suppose we collect a sample of size  $n$  of categorical data, consisting of  $r$  categories. This will sometimes be numerical data reduced to categories, as is done when constructing a histogram.

Suppose we hypothesize that these categories exists in the population according to frequencies  $p_1, \dots, p_r$ , and so we wish to test the hypotheses

$H_o$  :  $p_1, \dots, p_r$  are the population frequencies for categories  $1, \dots, r$

$H_a$  : At least one of the hypothetical frequencies is incorrect

The statistic we use is given by

$$X^2 = \sum_{i=1}^r \frac{(O_i - E_i)^2}{E_i}$$

where

$O_i$  = Observed count for category  $i = n_i$

$E_i$  = Expected count for category  $i = np_i$

The observed level of significance is

$$\alpha_{obs} = P(\chi_\nu^2 > X^2)$$

A size  $\alpha$  rejection region is given by

$$X^2 \geq \chi_{r-1,\alpha}^2.$$

When small cell sizes are present ( $\leq 5$ ) Yate's correction may be used in place of  $X^2$ :

$$X_{Yates}^2 = \sum_{i=1}^r \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

## D.13 Test for Independence in Contingency Tables

Suppose we are given a contingency table with  $n_r$  rows and  $n_c$  columns. Let

$r_i$  =  $P(i$ th row event occurs )

$c_j$  =  $P(j$ th column event occurs )

$p_{ij}$  =  $P(i$ th row event AND  $j$ th column event occur )

If the  $i$ th row event and the  $j$ th column event are independent then

$$p_{ij} = r_i c_j$$

so that the hypothesis of row/column independence can be written

$$\begin{aligned} H_o &: p_{ij} = r_i c_j \text{ for all row events } i \text{ and column events } j \\ H_a &: p_{ij} \neq r_i c_j \text{ for some row event } i \text{ and column event } j. \end{aligned}$$

The statistic we use is given by

$$X^2 = \sum_i^{n_r} \sum_j^{n_c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where

$$\begin{aligned} O_{ij} &= \text{Observed count in cell } i, j \\ E_{ij} &= \text{Expected count in cell } i, j \text{ under } H_o. \end{aligned}$$

The quantity  $O_{ij}$  is simply the count given in the cell given by row  $i$  and column  $j$ . The expected count is

$$E_{ij} = N r_i c_j,$$

where  $N$  is the total cell count. If we let

$$\begin{aligned} R_i &= \text{Total counts in row } i \\ C_j &= \text{Total counts in column } j \end{aligned}$$

then as an approximation we have

$$r_i \approx \frac{R_i}{N}, \quad c_j \approx \frac{C_j}{N}, \quad E_{ij} \approx N \frac{R_i}{N} \frac{C_j}{N} = \frac{R_i C_j}{N}.$$

This means the observed level of significance for rejecting  $H_o$  is

$$\alpha_{obs} = P(\chi_{\nu}^2 \geq X^2)$$

where  $\chi_{\nu}^2$  is a random variable with a  $\chi^2$  distribution with  $\nu = (n_r - 1)(n_c - 1)$  degrees of freedom. A size  $\alpha$  rejection region is given by

$$X^2 \geq \chi_{(n_r-1)(n_c-1), \alpha}^2.$$

When small cell sizes are present ( $\leq 5$ ) Yate's correction may be used in place of  $X^2$ :

$$X_{Yates}^2 = \sum_i^{n_r} \sum_j^{n_c} \frac{(|O_{ij} - E_{ij}| - 0.5)^2}{E_{ij}}.$$

## D.14 ANOVA

We often have situations in which we have  $k$  random samples from  $k$  distinct populations with population means  $\mu_1, \dots, \mu_k$ . Interest is then in testing the hypothesis

$$\begin{aligned} H_o &: \mu_1 = \mu_2 = \dots = \mu_k \\ H_a &: \mu_i \neq \mu_j \text{ for some } i, j \end{aligned}$$

In other words, are there some differences between the means ( $H_a$ ) or are they all the same ( $H_o$ ).

Pop'n	Pop'n Mean	Sample Size	Sample	Sample Mean	Sum of Squares	Sample Variance
1	$\mu_1$	$n_1$	$y_{11}, y_{12}, \dots, y_{1n_1}$	$\bar{y}_1$	$\sum_{i=1}^{n_1} (y_{1i} - \bar{y}_1)^2$	$S_1^2$
2	$\mu_2$	$n_2$	$y_{21}, y_{22}, \dots, y_{2n_2}$	$\bar{y}_2$	$\sum_{i=1}^{n_2} (y_{2i} - \bar{y}_2)^2$	$S_2^2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$k$	$\mu_k$	$n_k$	$y_{k1}, y_{k2}, \dots, y_{kn_k}$	$\bar{y}_k$	$\sum_{i=1}^{n_k} (y_{ki} - \bar{y}_k)^2$	$S_k^2$

The groups may be referred to as **treatments**. Sometimes it is convenient to refer to a treatment as a **factor** or **factor variable**. The observations  $y_{ij}$  are then **responses**, and  $\mu_i$  is a **mean response**. Here, we only have one factor, so the procedure is referred to as **one-way ANOVA**. If the sample sizes  $n_i$  are equal, we may refer to a **balanced design**.

We also have the **total mean**

$$\begin{aligned} \bar{y} &= \frac{\text{sum of all observations}}{n_1 + n_2 + \dots + n_k} \\ &= \frac{n_1 \bar{y}_1 + n_2 \bar{y}_2 + \dots + n_k \bar{y}_k}{n_1 + n_2 + \dots + n_k} \end{aligned}$$

In order to develop a test statistic for the hypothesis, we define the **treatment sum of squares**

$$\text{SST} = \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2$$

and the **error sum of squares**

$$\text{SSE} = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = \sum_{i=1}^k (n_i - 1) S_i^2$$

It can be shown that if define the **total sum of squares** to be

$$\text{SSTO} = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2.$$

then

$$\text{SSTO} = \text{SST} + \text{SSE}.$$

This is summarized in an ANOVA table:

Source	SS	df	MS	
Between Treatment (or Treatment)	SST	$k - 1$	$MST = \frac{\text{SST}}{k-1}$	$F = \frac{MST}{MSE}$
Within Treatment (or Error)	SSE	$n - k$	$MSE = \frac{\text{SSE}}{n-k}$	
Total	SSTot	$n - 1$		

The test statistic we use is then

$$F_{obs} = \frac{\text{SST}/(k - 1)}{\text{SSE}/(n - k)}$$

where

$$n = n_1 + n_2 + \dots + n_k.$$

To reject the null hypothesis we use the observed significance level defined by

$$\alpha_{obs} = P(F_{k-1,n-k} \geq F_{obs}),$$

or reject  $H_o$  with significance level  $\alpha$  if

$$F_{obs} \geq F_{k-1,n-k,\alpha}.$$

To compare  $m$  pairs of means with differences  $\mu_i - \mu_j$ , use the Bonferroni multiple comparison procedure, which gives confidence intervals:

$$\bar{y}_i - \bar{y}_j \pm t_{n-k,\alpha/(m^2)} \sqrt{MSE \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}.$$

#### D.14.1 Sign Test for Paired Comparisons

We are given a paired sample  $X_1, \dots, X_n, Y_1, \dots, Y_n$ , with differences  $D_i = X_i - Y_i$ ,  $i = 1, \dots, n$ . The differences can be represented as a random quantity  $D$ , and we have null hypothesis

$$H_o : \text{median of } D = 0.$$

We evaluate

$$\begin{aligned} T_+ &= \text{number of positive } D_i\text{'s,} \\ T_- &= \text{number of negative } D_i\text{'s.} \end{aligned}$$

If any of the differences  $D_i$  are zero, we discard them and adjust sample size  $n$  accordingly. To test

$$H_o : \text{median of } D = 0 \text{ against } H_a : \text{median of } D < 0$$

use

$$\alpha_{obs} = P(X \leq T_+)$$

where  $X \sim bin(n, 1/2)$ .

To test

$$H_o : \text{median of } D = 0 \text{ against } H_a : \text{median of } D > 0$$

use

$$\alpha_{obs} = P(X \leq T_-)$$

where  $X \sim bin(n, 1/2)$ .

To test

$$H_o : \text{median of } D = 0 \text{ against } H_a : \text{median of } D \neq 0$$

set

$$\alpha_{obs} = 2P(X \leq \min(T_-, T_+))$$

where  $X \sim bin(n, 1/2)$ .

## D.15 Wilcoxon Signed Rank Test for Paired Comparisons

We are given a paired sample  $X_1, \dots, X_n, Y_1, \dots, Y_n$ , with differences  $D_i = X_i - Y_i, i = 1, \dots, n$ . The differences can be represented as a random quantity  $D$ , and we have null hypothesis

$$H_o : \text{median of } D = 0.$$

For each value we calculate  $|D_i|$ , and then rank these values (use average ranks for ties). As in the sign test, we assign '+' or '-' according to whether the value is above or below the hypothetical media. We can designate each rank as *positive* or *negative* according to it's sign. We then set

$$\begin{aligned} T_+ &= \text{sum of positive ranks,} \\ T_- &= \text{sum of negative ranks,} \\ T_{obs} &= \min\{T_-, T_+\}. \end{aligned}$$

If any of the differences  $D_i$  are zero, we discard them and adjust sample size  $n$  accordingly. For large samples (say,  $n > 12$ ) we can use  $z$ -score

$$Z_{obs} = \frac{T_{obs} - \mu_T}{\sigma_T}$$

where

$$\mu_T = n(n+1)/4, \quad \text{and } \sigma_T = \sqrt{n(n+1)(2n+1)/24},$$

and use approximation  $Z_{obs} \sim N(0, 1)$ .

To test

$$H_o : \text{median of } D = 0 \text{ against } H_a : \text{median of } D < 0$$

reject  $H_o$  if  $T_{obs} = T_+$  with

$$\alpha_{obs} = P(Z \leq Z_{obs}).$$

To test

$$H_o : \text{median of } D = 0 \text{ against } H_a : \text{median of } D > 0$$

reject  $H_o$  if  $T_{obs} = T_-$  with

$$\alpha_{obs} = P(Z \leq Z_{obs}).$$

To test

$$H_o : \text{median of } D = 0 \text{ against } H_a : \text{median of } D \neq 0$$

set

$$\alpha_{obs} = 2P(Z \leq Z_{obs}).$$

## D.16 Wilcoxon Rank Sum Test for Independent Samples

We are given independent samples  $X_1, \dots, X_{n_1}$ ,  $Y_1, \dots, Y_{n_2}$ , and we have null hypothesis

$$H_o : \tilde{\mu}_1 = \tilde{\mu}_2$$

where  $\tilde{\mu}_1, \tilde{\mu}_2$  are the population medians. Both samples are pooled, then ranked (use average ranks for ties).

$$\begin{aligned} T_1 &= \text{sum of ranks from sample 1,} \\ T_2 &= \text{sum of ranks from sample 2.} \end{aligned}$$

Usually, samples are labeled (or can be relabeled) so that  $n_1 \leq n_2$ , and we only need to calculate  $T_1$ .

$$\mu_1 = n_1(n_1+n_2+1)/2, \quad \mu_2 = n_2(n_1+n_2+1)/2, \quad \text{and} \quad \sigma_W^2 = n_1n_2(n_1+n_2+1)/12.$$

$$Z_{obs} = \frac{T_1 - \mu_1}{\sigma_W}$$

**One-sided test (lower tail).** To test

$$H_o : \tilde{\mu}_1 \geq \tilde{\mu}_2 \text{ against } H_a : \tilde{\mu}_1 < \tilde{\mu}_2$$

use observed significance level

$$\alpha_{obs} = P(Z \leq Z_{obs}).$$

**One-sided test (upper tail).** To test

$$H_o : \tilde{\mu}_1 \leq \tilde{\mu}_2 \text{ against } H_a : \tilde{\mu}_1 > \tilde{\mu}_2$$

use observed significance level

$$\alpha_{obs} = P(Z \geq Z_{obs}).$$

**Two-sided test.** To test

$$H_o : \tilde{\mu}_1 = \tilde{\mu}_2 \text{ against } H_a : \tilde{\mu}_1 \neq \tilde{\mu}_2$$

use observed significance level

$$\alpha_{obs} = 2P(Z \leq -|Z_{obs}|).$$

# Bibliography

- Chambers, J. (1977). *Computational Methods for Data Analysis*. John Wiley & Sons.
- Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, **51**(4), 661–703.
- Devore, J. (1995). *Probability and Statistics for Engineering and the Sciences*. Duxbury Press, 4 edition.
- Diaconis, P., Holmes, S., and Montgomery, R. (2007). Dynamical bias in the coin toss. *SIAM Review*, **49**(2), 211–235.
- Feller, W. (1971). *Probability Theory and Its Applications, Volume 2*. John Wiley and Sons, New York, NY, second edition.
- Horne, A. D., Lachenbruch, P. A., Getson, P. R., and Hsu, H. S. (2001). Analysis of studies to evaluate immune response to combination vaccines. *Clinical Infectious Diseases*, **33**(Supplement 4), S306–S311.
- Hyndman, R. J. and Fan, Y. (1996). Sample quantiles in statistical packages. *The American Statistician*, **50**(4), 361–365.
- Karlin, S. and Taylor, H. M. (1975). *A First Course in Stochastic Processes*. Academic Press, San Diego, CA, second edition.
- Kendall, D. G. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded markov chain. *Annals of Mathematical Statistics*, **24**(3), 338–354.
- Kleinrock, L. (1975). *Queueing Systems (Volume 1: Theory)*. John Wiley and Sons, New York, NY.
- Ross, S. (2019). *Introduction to Probability Models (12th ed)*. Elsevier Science.
- Ross, S. M. (1996). *Stochastic Processes*. John Wiley and Sons, New York, NY, second edition.

# Index

- AUC*, 371  
*F* distribution, 304  
*P*-value, 215  
*iid* sample, 53  
*k*-step transition probability, 115  
*cumulative distribution function* (CDF), 37  
class intervals, 143  
density, 144  
interquartile range, 147  
lower fence, 148  
lower quartile, 146  
middle quartile, 146  
outliers, 148  
relative class frequencies, 144  
upper fence, 148  
upper quartile, 146  
whisker, 148  
absorbing state, 116  
accept, 209  
adjusted  $R^2$ , 340  
alternative hypothesis, 209  
analysis of variance, 303  
ANOVA, 303  
ANOVA summary table, 306  
area under curve, 371  
arrival process, 110  
average rank, 271  
balance equation, 122  
Bartlett's test, 265  
Bayes' theorem, 90  
bijective, 410  
binomial coefficient, 27  
binomial random variable, 61  
birth and death process, 123  
bootstrap procedure, 377  
Box-Cox power transformations, 174  
boxplot, 146  
cardinality, 14  
Categorical Data, 153  
Cauchy-Schwartz inequality, 56  
central limit theorem, 75  
central tendency, 133  
choose, 27  
codomain, 410  
coefficient of determination, 327  
coefficient of multiple determination, 338  
coefficient of variation, 163  
combination, 27  
complement, 13  
composite hypothesis, 209  
conditional, 20  
conditional odds, 93  
confidence interval, 208  
confidence level, 208  
contingency table, 21  
continuous-time Markov chain, 123  
correlation, 55  
correlation coefficient, 285  
Countable, 14  
counting process, 110  
covariance, 55  
critical value, 45  
De Morgan's Laws, 15

- degrees of freedom, 81  
denominator degrees of freedom, 82,  
    304  
density function, 34  
dependent variable, 320  
discrete, 14  
distribution, 33  
domain, 410  
  
empirical rule, 165, 166  
empty set, 14, 410  
error sum of squares, 304, 324  
expected value, 46  
exponential distribution, 68  
  
factorial, 26  
Finite, 14  
fitted values, 330  
floor function, 109  
full and reduced models, 338  
function, 410  
  
geometric random variable, 65  
gold standard, 99  
  
histogram, 143  
  
image, 410  
independent, 53  
independent variable, 320  
inference, 207  
infimum, 411  
injective, 410  
interaction term, 345  
intercept, 329  
intersection, 13  
inverse image, 411  
  
Kendall's notation, 127  
Kruskal-Wallis test, 310  
  
law of total probability, 23  
least squares estimate, 337  
level of significance, 210  
  
likelihood ratio, 93  
linear congruential generator, 103  
linear least squares coefficients, 321  
linear transformation, 52  
  
main effects, 345  
Markov chain, 112  
maximum, 411  
mean error sum of squares, 325  
memoryless property, 66, 112  
minimum, 411  
model selection, 339  
Monte Carlo method, 102  
multinomial coefficient, 30  
multiple linear regression, 320  
multiple regression, 336  
mutually exclusive, 15  
  
nested models, 339  
new better than used (NBU), 84  
new worse than used (NWU), 84  
Nominal Data, 153  
normal distribution, 69  
normalization, 35  
nuisance parameter, 207  
nul hypothesis, 209  
numerator degrees of freedom, 82,  
    304  
  
observed level of significance, 215  
odds, 92  
odds ratio, 101  
one-to-one, 410  
onto, 410  
order statistics, 134  
ordered selection, 25  
Ordinal Data, 153  
  
P-value, 199  
pairwise independent, 20  
pairwise multiple comparisons., 309  
parameter space, 207  
parametric family of distributions,  
    207

- Pearson product-moment correlation coefficient, 285  
percentile, 43  
permutation, 27  
permutation test, 374  
pilot study, 247  
Poisson process, 110  
polynomial regression, 360  
posterior odds, 93  
posterior probability, 90  
predicted responses, 322  
predicted value, 323  
prediction bounds, 332  
predictor, 320  
preimage, 411  
prior odds, 93  
prior probability, 90  
probability mass function, 34  
prosecutor's fallacy, 96  
quadratic terms, 360  
quantile, 43  
quantile function, 43  
quantile plot, 167  
random variable, 33  
range, 410  
receiver operating characteristic, 370  
regression sum of squares, 326  
reject, 209  
rejection region, 210  
renewal process, 110  
reproducibility, 102  
residuals, 323  
risk score, 364  
ROC curve, 370  
rule of product, 25  
sample average, 58  
sample correlation coefficient, 187  
sample covariance, 187  
sample quantile, 137  
sample space, 13  
sample standard deviation, 160  
sample variance, 158  
scatter plot, 185  
seed, 103  
simple hypothesis, 209  
simple linear regression, 320  
size, 210  
skewness, 163  
slope, 329  
Spearman rank correlation, 288  
standard deviation, 51  
standard errors, 331  
state space, 112  
statistical model, 342  
steady state frequency, 121  
stochastic matrix, 113  
stochastic process, 110  
support, 34  
supremum, 411  
surjective, 410  
survival time, 83  
tail probability, 38  
total mean, 304  
total sum of squares, 304  
transformation, 48  
transition probability, 112  
transition probability matrix, 112  
transpose the conditional, 95  
treatment sum of squares, 304, 326  
Tukey's pairwise procedure, 309  
Tukey-Kramer pairwise procedure, 309  
type I error, 213  
type II error, 213  
Uncountable, 14  
uniform distribution, 67  
union, 13  
unordered selection, 25  
upper confidence bound, 261  
utilization factor, 126  
variance, 51

Venn diagram, 17

Yates's correction for continuity, 296