

PROBABILITY *and* STATISTICS WITH

R

María Dolores Ugarte
Ana F. Militino
Alan T. Arnholt

 CRC Press
Taylor & Francis Group

A CHAPMAN & HALL BOOK

PROBABILITY *and*
STATISTICS WITH *R*

PROBABILITY *and* STATISTICS WITH *R*

María Dolores Ugarte

Ana F. Militino

Alan T. Arnholt



CRC Press

Taylor & Francis Group

Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2008 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works
Version Date: 20140904

International Standard Book Number-13: 978-1-58488-892-5 (eBook - PDF)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (<http://www.copyright.com/>) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
<http://www.taylorandfrancis.com>

and the CRC Press Web site at
<http://www.crcpress.com>

Preface

The authors would like to thank their parents

Lola: Pedro and Loli

Ana: Carmelo and Juanita

Alan: Terry and Loretta

for their unflagging support and encouragement.

The Book

Probability and Statistics with R is a work born of the love of statistics and the advancements that have been made in the field as more powerful computers can be used to perform calculations and simulations that were only dreamed of by those who came before. The S language and its derivative, R, have made the practice of statistics available to anyone with the time and inclination to do so.

Teachers will enjoy the real-world examples and the thoroughly worked out derivations. Those wanting to use this book as a reference work will appreciate the extensive treatments on data analysis using appropriate techniques, both parametric and nonparametric. Students who are visual learners will appreciate the detailed graphics and clear captions, while the hands-on learners will be pleased with the abundant problems and solutions. (A solutions manual should be available from Taylor & Francis.) It is our hope that practitioners of statistics at every level will welcome the features of this book and that it will become a valuable addition to their statistics libraries.

The Purpose

Our primary intention when we undertook this project was to introduce R as a teaching statistical package, rather than just a program for researchers. As much as possible, we have made a great effort to link the statistical contents with the procedures used by R to show consistency to undergraduate students. The reader who uses S-PLUS will also find this text useful, as S-PLUS commands are included with those for R in the vast majority of the examples.

This book is intended to be practical, readable, and clear. It gives the reader real-world examples of how S can be used to solve problems in every topic covered including, but not limited to, general probability in both the univariate and multivariate cases, sampling distributions and point estimation, confidence intervals, hypothesis testing, experimental design, and regression. Most of the problems are taken from genuine data sets rather than created out of thin air. Next, it is unusually thorough in its treatment of virtually every topic, covering both the traditional methods to solve problems as well as many nonparametric techniques. Third, the figures used to explain difficult topics are exceptionally detailed.

Finally, the derivations of difficult equations are worked out thoroughly rather than being left as exercises. These features, and many others, will make this book beneficial to any reader interested in applying the S language to the world of statistics.

The Program

The S language includes both R and S-PLUS. “R can be regarded as an implementation of the S language which was developed at Bell Laboratories by Rick Becker, John Chambers, and Allan Wilks, and also forms the basis of the S-PLUS systems.”

(<http://cran.r-project.org/doc/manuals/R-intro.html#Preface>)

The current R is the result of a collaborative effort with contributions from all over the world. R was initially written by Robert Gentleman and Ross Ihaka of the Statistics Department of the University of Auckland. Since mid-1997 there has been a core group with write access to the R source (<http://www.r-project.org/>—click “Contributors” on the sidebar).

Not only is R an outstanding statistical package, but it is offered free of charge and can be downloaded from <http://www.r-project.org/>. The authors are greatly indebted to the giants of statistics and programming on whose shoulders we have stood to see what we will show the readers of this text.

The Content

The core of the material covered in this text has been used in undergraduate courses at the Public University of Navarre for the last ten years. It has been used to teach engineering (agricultural, industrial, and telecommunications) and economics majors. Some of the material in this book has also been used to teach graduate students studying agriculture, biology, engineering, and medicine.

The book starts with a brief introduction to S that includes syntax, structures, and functions. It is designed to provide an overview of how to use both R and S-PLUS so that even a neophyte will be able to solve the problems by the end of the chapter.

Chapter 2, entitled “Exploring Data,” covers important graphical and numerical descriptive methods. This chapter could be used to teach a first course in statistics.

The next three chapters deal with probability and random variables in a generally classical presentation that includes many examples and an extensive collection of problems to practice all that has been learned.

Chapter 6 presents some important statistics and their sampling distributions. Solving the exercises will give any reader confidence that the difficult topics covered in this chapter are understood.

The next four chapters encompass point estimation, confidence intervals, hypothesis testing, and a wide range of nonparametric methods including goodness-of-fit tests, categorical data analysis, nonparametric bootstrapping, and permutation tests.

Chapter 11 provides an introduction to experimental design using fixed and random effects models as well as the randomized block design and the two-factor factorial design.

The book ends with a chapter on simple and multiple regression analysis. The procedures from this chapter are used to solve three interesting case studies based on real data.

The Fonts

Knowing several typographical conventions will help the reader in understanding the material presented in this text. R code is displayed in a monospaced font with the > symbol in front of commands that are entered at the R prompt.

```
> x<-0.28354
> round(x,2)
[1] 0.28
```

The same font is used for data sets and functions, though functions are followed by (). For example, the `PASWR` package but the `round()` function would be shown. Throughout the text, a ■ is found at the end of solutions to examples. In the index, page numbers in **BOLD** are where the primary occurrences of topics are found, while those in *ITALICS* indicate the pages where a problem about a topic or using a given data set can be located.

The Web

This text is supported at <http://www1.appstate.edu/~arnholta/PASWR> on the Internet. The website has up-to-date errata, chapter scripts, and a copy of the `PASWR` package (which is also on CRAN) available for download.

Acknowledgments

We gratefully acknowledge the invaluable help provided by Susie Arnholt. Her willingness to apply her expertise in L^AT_EX and knowledge of English grammar to the production of this book is appreciated beyond words.

Several people were instrumental in improving the overall readability of this text. The recommendations made by Phil Spector, the Applications Manager and Software Consultant for the Statistical Computing Facility in the Department of Statistics at the University of California at Berkeley, who reviewed this text for Taylor & Francis, were used in improving much of the original R code as well as decreasing the inevitable typographical error rate. Tomás Goicoa, a member of the Spatial Statistics Research Group at the Public University of Navarre, was of great help in preparing and checking exercises. Celes Alexander, an Appalachian State University graduate student, graciously read the entire text and found several typos. Any remaining typos or errors are entirely the fault of the authors.

Thanks to our editor at Taylor & Francis, David Grubbs, for embracing and encouraging our project. Many thanks the Statistics and Operations Research Department at Public University of Navarre and to the Department of Mathematical Sciences at Appalachian State University for the support they gave us in our writing of this text.

The “You choose, you decide” initiative sponsored by Caja Navarra also provided funding for in-person collaborations. Thanks to the *Universidad Nacional de Educación a Distancia*, in particular the *Centro Asociado de Pamplona*, for allowing us to present this project under their auspices.

Special thanks to José Luis Iriarte, the former Vicerector of International Relations of the Public University of Navarre, and to T. Marvin Williamsen, the former Associate Vice Chancellor for International Programs at Appalachian State University. These men were instrumental in gaining funding and support for several in-person collaborations including a year-long visit at the Public University of Navarre for the third author and two multi-week visits for the first two authors to Appalachian State University.

Finally, to the geniuses of this age who first conceived of the idea of an excellent open source software for statistics and those who reared the idea to adulthood, our gratitude is immeasurable. May the lighthouse of your brilliance guide travelers on the ocean of statistics for decades to come. Thank you, R Core Team.

Contents

1	A Brief Introduction to S	1
1.1	The Basics of S	1
1.2	Using S	1
1.3	Data Sets	2
1.4	Data Manipulation	3
1.4.1	S Structures	3
1.4.2	Mathematical Operations	4
1.4.3	Vectors	4
1.4.4	Sequences	5
1.4.5	Reading Data	7
1.4.5.1	Using <code>scan()</code>	7
1.4.5.2	Using <code>read.table()</code>	8
1.4.5.3	Using <code>write()</code>	8
1.4.5.4	Using <code>dump()</code> and <code>source()</code>	9
1.4.6	Logical Operators and Missing Values	9
1.4.7	Matrices	12
1.4.8	Vector and Matrix Operations	14
1.4.9	Arrays	15
1.4.10	Lists	16
1.4.11	Data Frames	16
1.4.12	Tables	17
1.4.13	Functions Operating on Factors and Lists	19
1.5	Probability Functions	20
1.6	Creating Functions	21
1.7	Programming Statements	22
1.8	Graphs	23
1.9	Problems	25
2	Exploring Data	29
2.1	What Is Statistics?	29
2.2	Data	29
2.3	Displaying Qualitative Data	30
2.3.1	Tables	30
2.3.2	Barplots	31
2.3.3	Dot Charts	32
2.3.4	Pie Charts	32
2.4	Displaying Quantitative Data	33
2.4.1	Stem-and-Leaf Plots	33
2.4.2	Strip Charts (R Only)	35
2.4.3	Histograms	36
2.5	Summary Measures of Location	39
2.5.1	The Mean	39
2.5.2	The Median	41

2.5.3	Quantiles	42
2.5.4	Hinges and Five-Number Summary	44
2.5.5	Boxplots	45
2.6	Summary Measures of Spread	47
2.6.1	Range	47
2.6.2	Interquartile Range	47
2.6.3	Variance	48
2.7	Bivariate Data	49
2.7.1	Two-Way Contingency Tables	49
2.7.2	Graphical Representations of Two-Way Contingency Tables	51
2.7.3	Comparing Samples	53
2.7.4	Relationships between Two Numeric Variables	56
2.7.5	Correlation	58
2.7.6	Sorting a Data Frame by One or More of Its Columns	59
2.7.7	Fitting Lines to Bivariate Data	60
2.8	Multivariate Data (Lattice and Trellis Graphs)	65
2.8.1	Arranging Several Graphs on a Single Page	67
2.8.2	Panel Functions	69
2.9	Problems	71
3	General Probability and Random Variables	77
3.1	Introduction	77
3.2	Counting Rules	77
3.2.1	Sampling With Replacement	77
3.2.2	Sampling Without Replacement	78
3.2.3	Combinations	79
3.3	Probability	80
3.3.1	Sample Space and Events	80
3.3.2	Set Theory	80
3.3.3	Interpreting Probability	81
3.3.3.1	Relative Frequency Approach to Probability	81
3.3.3.2	Axiomatic Approach to Probability	81
3.3.4	Conditional Probability	83
3.3.5	The Law of Total Probability and Bayes' Rule	84
3.3.6	Independent Events	86
3.4	Random Variables	87
3.4.1	Discrete Random Variables	88
3.4.2	Mode, Median, and Percentiles	89
3.4.3	Expected Values of Discrete Random Variables	90
3.4.4	Moments	92
3.4.4.1	Variance	92
3.4.4.2	Rules of Variance	92
3.4.5	Continuous Random Variables	93
3.4.5.1	Numerical Integration with S	96
3.4.5.2	Mode, Median, and Percentiles	96
3.4.5.3	Expectation of Continuous Random Variables	98
3.4.6	Markov's Theorem and Chebyshev's Inequality	100
3.4.7	Weak Law of Large Numbers	102
3.4.8	Skewness	102
3.4.9	Moment Generating Functions	104
3.5	Problems	107

4	Univariate Probability Distributions	115
4.1	Introduction	115
4.2	Discrete Univariate Distributions	115
4.2.1	Discrete Uniform Distribution	115
4.2.2	Bernoulli and Binomial Distributions	116
4.2.3	Poisson Distribution	120
4.2.4	Geometric Distribution	126
4.2.5	Negative Binomial Distribution	128
4.2.6	Hypergeometric Distribution	129
4.3	Continuous Univariate Distributions	130
4.3.1	Uniform Distribution (Continuous)	130
4.3.2	Exponential Distribution	133
4.3.3	Gamma Distribution	139
4.3.4	Hazard Function, Reliability Function, and Failure Rate	143
4.3.5	Weibull Distribution	147
4.3.6	Beta Distribution	149
4.3.7	Normal (Gaussian) Distribution	152
4.4	Problems	162
5	Multivariate Probability Distributions	171
5.1	Joint Distribution of Two Random Variables	171
5.1.1	Joint pdf for Two Discrete Random Variables	171
5.1.2	Joint pdf for Two Continuous Random Variables	173
5.2	Independent Random Variables	174
5.3	Several Random Variables	175
5.4	Conditional Distributions	177
5.5	Expected Values, Covariance, and Correlation	180
5.5.1	Expected Values	180
5.5.2	Covariance	181
5.5.3	Correlation	183
5.6	Multinomial Distribution	185
5.7	Bivariate Normal Distribution	186
5.8	Problems	190
6	Sampling and Sampling Distributions	197
6.1	Sampling	197
6.1.1	Simple Random Sampling	198
6.1.2	Stratified Sampling	200
6.1.3	Systematic Sampling	200
6.1.4	Cluster Sampling	201
6.2	Parameters	201
6.2.1	Infinite Populations' Parameters	202
6.2.2	Finite Populations' Parameters	202
6.3	Estimators	203
6.3.1	Empirical Probability Distribution Function	204
6.3.2	Plug-In Principle	206
6.4	Sampling Distribution of the Sample Mean	206
6.5	Sampling Distribution for a Statistic from an Infinite Population	212
6.5.1	Sampling Distribution for the Sample Mean	212
6.5.1.1	First Case: Sampling Distribution of \bar{X} when Sampling from a Normal Distribution	212

6.5.1.2	Second Case: Sampling Distribution of \bar{X} when X Is not a Normal Random Variable	215
6.5.2	Sampling Distribution for $\bar{X} - \bar{Y}$ when Sampling from Two Independent Normal Populations	219
6.5.3	Sampling Distribution for the Sample Proportion	220
6.5.4	Expected Value and Variance of the Uncorrected Sample Variance and the Sample Variance	225
6.6	Sampling Distributions Associated with the Normal Distribution	226
6.6.1	Chi-Square Distribution (χ^2)	226
6.6.1.1	The Relationship between the χ^2 Distribution and the Normal Distribution	228
6.6.1.2	Sampling Distribution for S_u^2 and S^2 when Sampling from Normal Populations	231
6.6.2	t -Distribution	235
6.6.3	The F Distribution	238
6.7	Problems	241
7	Point Estimation	245
7.1	Introduction	245
7.2	Properties of Point Estimators	245
7.2.1	Mean Square Error	245
7.2.2	Unbiased Estimators	247
7.2.3	Efficiency	249
7.2.4	Consistent Estimators	252
7.2.5	Robust Estimators	254
7.3	Point Estimation Techniques	255
7.3.1	Method of Moments Estimators	255
7.3.2	Likelihood and Maximum Likelihood Estimators	257
7.3.2.1	Fisher Information	270
7.3.2.2	Fisher Information for Several Parameters	271
7.3.2.3	Properties of Maximum Likelihood Estimators	273
7.3.2.4	Finding Maximum Likelihood Estimators for Multiple Parameters	278
7.3.2.5	Multi-Parameter Properties of MLEs	280
7.4	Problems	282
8	Confidence Intervals	291
8.1	Introduction	291
8.2	Confidence Intervals for Population Means	292
8.2.1	Confidence Interval for the Population Mean when Sampling from a Normal Distribution with Known Population Variance	292
8.2.1.1	Determining Required Sample Size	297
8.2.2	Confidence Interval for the Population Mean when Sampling from a Normal Distribution with Unknown Population Variance	300
8.2.3	Confidence Interval for the Difference in Population Means when Sampling from Independent Normal Distributions with Known Equal Variances	302
8.2.4	Confidence Interval for the Difference in Population Means when Sampling from Independent Normal Distributions with Known but Unequal Variances	305

8.2.5	Confidence Interval for the Difference in Means when Sampling from Independent Normal Distributions with Variances That Are Unknown but Assumed Equal	308
8.2.6	Confidence Interval for a Difference in Means when Sampling from Independent Normal Distributions with Variances That Are Unknown and Unequal	310
8.2.7	Confidence Interval for the Mean Difference when the Differences Have a Normal Distribution	313
8.3	Confidence Intervals for Population Variances	316
8.3.1	Confidence Interval for the Population Variance of a Normal Population	316
8.3.2	Confidence Interval for the Ratio of Population Variances when Sampling from Independent Normal Distributions	319
8.4	Confidence Intervals Based on Large Samples	321
8.4.1	Confidence Interval for the Population Proportion	322
8.4.2	Confidence Interval for a Difference in Population Proportions	327
8.4.3	Confidence Interval for the Mean of a Poisson Random Variable	329
8.5	Problems	331
9	Hypothesis Testing	341
9.1	Introduction	341
9.2	Type I and Type II Errors	342
9.3	Power Function	345
9.4	Uniformly Most Powerful Test	348
9.5	ϕ -Value or Critical Level	350
9.6	Tests of Significance	351
9.7	Hypothesis Tests for Population Means	353
9.7.1	Test for the Population Mean when Sampling from a Normal Distribution with Known Population Variance	353
9.7.2	Test for the Population Mean when Sampling from a Normal Distribution with Unknown Population Variance	355
9.7.3	Test for the Difference in Population Means when Sampling from Independent Normal Distributions with Known Variances	361
9.7.4	Test for the Difference in Means when Sampling from Independent Normal Distributions with Variances That Are Unknown but Assumed Equal	363
9.7.5	Test for a Difference in Means when Sampling from Independent Normal Distributions with Variances That Are Unknown and Unequal	367
9.7.6	Test for the Mean Difference when the Differences Have a Normal Distribution	370
9.8	Hypothesis Tests for Population Variances	373
9.8.1	Test for the Population Variance when Sampling from a Normal Distribution	373
9.8.2	Test for Equality of Variances when Sampling from Independent Normal Distributions	376
9.9	Hypothesis Tests for Population Proportions	379
9.9.1	Testing the Proportion of Successes in a Binomial Experiment (Exact Test)	379
9.9.2	Testing the Proportion of Successes in a Binomial Experiment (Normal Approximation)	383

9.9.3	Testing Equality of Proportions with Fisher's Exact Test	387
9.9.4	Large Sample Approximation for Testing the Difference of Two Proportions	392
9.10	Problems	396
10	Nonparametric Methods	403
10.1	Introduction	403
10.2	Sign Test	403
10.2.1	Confidence Interval Based on the Sign Test	404
10.2.2	Normal Approximation to the Sign Test	405
10.3	Wilcoxon Signed-Rank Test	410
10.3.1	Confidence Interval for ψ Based on the Wilcoxon Signed-Rank Test	414
10.3.2	Normal Approximation to the Wilcoxon Signed-Rank Test	418
10.4	The Wilcoxon Rank-Sum or the Mann-Whitney U -Test	423
10.4.1	Confidence Interval Based on the Mann-Whitney U -Test	427
10.4.2	Normal Approximation to the Wilcoxon Rank-Sum and Mann-Whitney U -Tests	429
10.5	The Kruskal-Wallis Test	436
10.6	Friedman Test for Randomized Block Designs	442
10.7	Goodness-of-Fit Tests	447
10.7.1	The Chi-Square Goodness-of-Fit Test	447
10.7.2	Kolmogorov-Smirnov Goodness-of-Fit Test	454
10.7.3	Shapiro-Wilk Normality Test	461
10.8	Categorical Data Analysis	462
10.8.1	Test of Independence	464
10.8.2	Test of Homogeneity	466
10.9	Nonparametric Bootstrapping	469
10.9.1	Bootstrap Paradigm	469
10.9.2	Confidence Intervals	472
10.10	Permutation Tests	479
10.11	Problems	484
11	Experimental Design	491
11.1	Introduction	491
11.2	Fixed Effects Model	495
11.3	Analysis of Variance (ANOVA) for the One-Way Fixed Effects Model	497
11.4	Power and the Non-Central F Distribution	501
11.5	Checking Assumptions	510
11.5.1	Checking for Independence of Errors	510
11.5.2	Checking for Normality of Errors	511
11.5.3	Checking for Constant Variance	512
11.6	Fixing Problems	514
11.6.1	Non-Normality	515
11.6.2	Non-Constant Variance	516
11.7	Multiple Comparisons of Means	518
11.7.1	Fisher's Least Significant Difference	519
11.7.2	The Tukey's Honestly Significant Difference	520
11.7.3	Displaying Pairwise Comparisons	521
11.8	Other Comparisons among the Means	522
11.8.1	Orthogonal Contrasts	523
11.8.2	The Scheffé Method for All Contrasts	529

11.9	Summary of Comparisons of Means	529
11.10	Random Effects Model (Variance Components Model)	534
11.11	Randomized Complete Block Design	537
11.12	Two-Factor Factorial Design	547
11.13	Problems	556
12	Regression	563
12.1	Introduction	563
12.2	Simple Linear Regression	565
12.3	Multiple Linear Regression	565
12.4	Ordinary Least Squares	567
12.5	Properties of the Fitted Regression Line	570
12.6	Using Matrix Notation with Ordinary Least Squares	571
12.7	The Method of Maximum Likelihood	576
12.8	The Sampling Distribution of $\hat{\beta}$	577
12.9	ANOVA Approach to Regression	580
12.9.1	ANOVA with Simple Linear Regression	581
12.9.2	ANOVA with Multiple Linear Regression	584
12.9.3	Coefficient of Determination	586
12.9.4	Extra Sum of Squares	587
12.9.4.1	Tests on a Single Parameter	589
12.9.4.2	Tests on Subsets of the Regression Parameters	591
12.10	General Linear Hypothesis	593
12.11	Model Selection and Validation	597
12.11.1	Testing-Based Procedures	597
12.11.1.1	Backward Elimination	597
12.11.1.2	Forward Selection	597
12.11.1.3	Stepwise Regression	598
12.11.1.4	Criterion-Based Procedures	598
12.11.1.5	Summary	606
12.11.2	Diagnostics	607
12.11.2.1	Checking Error Assumptions	607
12.11.2.1.1	Assessing Normality and Constant Variance	608
12.11.2.1.2	Testing Autocorrelation	609
12.11.2.2	Identifying Unusual Observations	610
12.11.2.3	High Leverage Observations	613
12.11.3	Transformations	620
12.11.3.1	Collinearity	623
12.11.3.2	Transformations for Non-Normality and Unequal Error Variances	626
12.12	Interpreting a Logarithmically Transformed Model	630
12.13	Qualitative Predictors	632
12.14	Estimation of the Mean Response for New Values \mathbf{X}_h	638
12.15	Prediction and Sampling Distribution of New Observations $Y_{h(\text{new})}$	639
12.16	Simultaneous Confidence Intervals	642
12.16.1	Simultaneous Confidence Intervals for Several Mean Responses — Confidence Band	642
12.16.2	Predictions of g New Observations	643
12.16.3	Distinguishing Pointwise Confidence Envelopes from Confidence Bands	643
12.17	Problems	648

A S Commands	659
B Quadratic Forms and Random Vectors and Matrices	671
B.1 Quadratic Forms	671
B.2 Random Vectors and Matrices	672
B.3 Variance of Random Vectors	672
References	675
Index	683

List of Figures

1.1	Structures in <code>S</code>	3
1.2	Examples of the <code>plot()</code> function	24
1.3	Size, color, and choice of plotting symbol	24
1.4	Autonomous communities in Spain	26
2.1	Graphical representation of the data in <code>Grades</code> and <code>Age</code> with the function <code>barplot()</code>	31
2.2	Graphical representation of the data in <code>Grades</code> and <code>Age</code> with the function <code>dotchart()</code>	32
2.3	Graphical representation of the data in <code>Grades</code> and <code>Age</code> with the function <code>pie()</code>	33
2.4	Nine different graphs labeled according to their shape	34
2.5	Strip chart of the number of home runs Babe Ruth hit while playing for the New York Yankees	36
2.6	Strip chart of the number of home runs Babe Ruth hit per season according to the team for which he was playing	36
2.7	Histograms created using different bin definitions for the number of home runs hit by Babe Ruth while playing for the New York Yankees	37
2.8	Histograms created using different bin definitions for the eruption duration of Old Faithful	38
2.9	Histogram of waiting time between Old Faithful eruptions with superimposed density estimate as well as a density plot	39
2.10	Boxplot illustration	46
2.11	Boxplot of car prices with five-number summaries labeled	47
2.12	Stacked and side-by-side barplots for levels of palpitation (<code>Teasy</code>) and physician (<code>Doctor</code>)	52
2.13	Side-by-side barplots showing percentages in obstructive contacts categories by treatments	53
2.14	Side-by-side barplots showing percentages in obstructive contacts' categories by treatments	55
2.15	Side-by-side boxplots of BWI in the traditional sitting and hamstring stretch positions	56
2.16	Density plots of BWI in the traditional sitting and hamstring stretch positions	56
2.17	Quantile-quantile plot of BWI in the traditional sitting and hamstring stretch positions	57
2.18	Scatterplot of $\log(\text{brain})$ versus $\log(\text{body})$ for Example 2.21	58
2.19	Graph depicting residuals. The vertical distances shown with a dotted line between the Y_i s, depicted with a solid circle, and the \hat{Y}_i s, depicted with a clear square, are the residuals.	62
2.20	Scatterplot of $\log(\text{brain})$ versus $\log(\text{body})$ with superimposed regression lines computed with (solid line) and without (dashed line) dinosaurs	63

2.21	Scatterplot of $\log(\mathbf{brain})$ versus $\log(\mathbf{body})$ with three superimposed regression lines. Solid is the OLS line; dashed is the least-trimmed squares line; and dotted is the robust line.	65
2.22	Comparative histograms of BWI by treatment	66
2.23	Trellis side-by-side boxplots of BWI in the traditional sitting and hamstring stretch positions given <code>Doctor</code>	67
2.24	Trellis side-by-side stripplots of BWI in the traditional sitting and hamstring stretch positions given <code>Doctor</code>	68
2.25	Arrangement of four different Trellis graphs on the same page	69
2.26	x - y plot of height (cm) versus weight (kg) given physician (<code>Doctor</code>) with superimposed least squares and least-trimmed squares lines	70
3.1	Sample space for car batteries example	85
3.2	Circuit system diagram	87
3.3	The pdf and cdf for coin tossing	90
3.4	Fulcrum illustration of $E[X]$	90
3.5	Illustration of $\mathbb{P}(a \leq X \leq b) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a)$	93
3.6	Illustration of pdf and cdf for calculations example	95
3.7	Graph of $2 \cos(2x)$ from 0 to $\frac{\pi}{4}$ with R	98
3.8	Skewness Illustrations	103
4.1	$Bin(0.3, 8)$ pdf and cdf	118
4.2	Comparison of simulated and theoretical binomial distributions	119
4.3	$Pois(1)$ pdf and cdf	124
4.4	The pdf and cdf for the random variable $X \sim Unif(a, b)$	131
4.5	The pdf and cdf for the random variable $X \sim Exp(\lambda = 0.75)$	135
4.6	Histogram of time between goals with superimposed exponential density curve	139
4.7	Graphical illustration of Γ random variables	141
4.8	Hazard functions with pdfs	144
4.9	Hazard function for printer failure	147
4.10	Weibull distributions	148
4.11	β distributions	150
4.12	Normal distributions with increasing σ values	153
4.13	Graphical representation for computing $\mathbb{P}(a \leq X \leq b)$	154
4.14	Graphical representation for finding $\mathbb{P}(90 \leq X \leq 115)$ given $X \sim N(100, 10)$	156
4.15	Quantile-quantile plot of the standardized test scores of 20 randomly selected college freshmen	159
4.16	Superimposed quantile-quantile plots	160
4.17	Resulting quantile-quantile plots using the function <code>ntester()</code> on standardized test scores	161
4.18	Graphical results from <code>EDA(scores)</code>	161
5.1	Graphical representation of the domain of interest for joint PDF example	174
5.2	Graphical representation of $f_{X,Y}(x, y) = 8xy, 0 \leq y \leq x \leq 1$	179
5.3	Scatterplots showing positive, negative, and zero covariance between two random variables where $p_{X,Y}(x, y) = 1/10$ for each of the ten pairs of plotted points.	182
5.4	Bivariate normal density representations	188
6.1	Empirical cumulative distribution function of rolling a die 100 times	206

6.2	Sampling distributions of \bar{X} and S^2 under random sampling (RS) and simple random sampling (SRS)	212
6.3	Comparison of uniform and normal graphs	215
6.4	Simulations of uniform and exponential distributions	216
6.5	Uniform and exponential simulations	217
6.6	Probability histogram with normal density	221
6.7	Illustrations of the pdfs of χ_3^2 , χ_6^2 , and χ_{16}^2 random variables	227
6.8	Probability histograms for simulated distributions of $(n - 1)S^2/\sigma^2$ when sampling from normal and exponential distributions	236
6.9	Illustrations of the pdfs of t_1 (dashed line), t_3 (dotted line), and t_∞ (solid line) random variables.	237
6.10	Illustrations of the pdfs of $F_{2,4}$ (solid line), $F_{4,9}$ (dotted line), and $F_{19,19}$ (dashed line) random variables	239
7.1	Visual representations of variance and bias	246
7.2	Graphical representations for the sampling distributions of $\hat{\mu}_1$ and $\hat{\mu}_2$	252
7.3	Illustration of the $\ln L(\pi \mathbf{x})$ function for the Oriental Cockroaches Example	260
7.4	Illustration of the $\ln L(p \mathbf{x})$ function for a General MLE Example	264
7.5	Illustration of the likelihood function in the I.I.D. Uniform Random Variables Example	266
7.6	Illustration of $\ln L(\mu \mathbf{x})$ and $\ln L(\sigma^2 \mathbf{x})$	269
8.1	Standard normal distribution with an area of $\alpha/2$ in each tail	293
8.2	Simulated confidence intervals for the population mean when sampling from a normal distribution with known variance	295
8.3	Quantile-quantile (normal distribution) plot of weekly monies spent on groceries for 30 randomly selected Watauga households	296
8.4	Quantile-quantile plot of the asking price for 14 randomly selected three-bedroom/two-bath houses in Watauga County, North Carolina	302
8.5	Superimposed normal quantile-quantile plots of the hardness values for fresh and warehoused apples	304
8.6	Superimposed normal quantile-quantile plots of mathematical assessment scores	307
8.7	Normal quantile-quantile plot of the time differences between Sun and Digital workstations to complete complex simulations	315
8.8	Quantile-quantile plot of the time differences between Sun and Digital workstations to complete complex simulations shown in the middle with normal quantile-quantile plots of random normal data depicted on the outside plots	315
8.9	Chi-square distribution with six degrees of freedom depicting the points $\chi_{\alpha/2;6}^2$ and $\chi_{1-\alpha/2;6}^2$	316
8.10	Quantile-quantile plot of 1932 barley yield in bushels/acre	318
8.11	F distribution with ten and ten degrees of freedom depicting the points $f_{\alpha/2;10,10}$ and $f_{1-\alpha/2;10,10}$	320
9.1	Graphical representation of type I and type II errors when $H_0 : \mu = 1$ versus $H_1 : \mu = 4$	345
9.2	Graphical representation of the power function, $Power(\mu)$, for both scenarios in the Achievement Test Example	348
9.3	Graphical representation of type I and type II errors when $H_0 : \mu = 1$ versus $H_1 : \mu = 2$ with rejection region $(2.036, \infty)$	349

9.4	Graphical representation of type I and type II errors when $H_0 : \mu = 1$ versus $H_1 : \mu = 2$ with rejection region $(1.100, 1.300) \cup (2.461, \infty)$	350
9.5	Central t -distribution and non-central t -distribution with $\gamma = 3$	358
9.6	Central t -distribution and simulated non-central t -distribution with $\gamma = 3$	359
9.7	Exploratory data analysis of the wheat yield per plot values	360
9.8	Side-by-side boxplots and normal quantile-quantile plots of the satisfaction level for graduates from State School X and State School Y	365
9.9	Side-by-side boxplots and normal quantile-quantile plots of the sodium content for source X and source Y.	369
9.10	Exploratory data analysis of the differences between 1932 barley yields from the Morris and Crookston sites.	372
9.11	Graphs from using <code>EDA()</code> on the washers' diameters	375
9.12	Exploratory data analysis for the blood alcohol values using the breathalyzers from company X and company Y on two volunteers after drinking four beers.	377
10.1	Graphical representation of a $Bin(20, 0.5)$ distribution and a superimposed normal distribution with $\mu_S = 20(0.5) = 10$ and $\sigma_S = \sqrt{20(0.5)^2} = 3.16$	406
10.2	Graphical representation of the data in <code>call.time</code> with the function <code>EDA()</code>	408
10.3	Horizontal boxplot of bus waiting times in minutes	416
10.4	Graphical representation of the Wilcoxon signed-rank distribution for $n = 15$ superimposed by a normal distribution with $\mu = n(n + 1)/4 = 60$ and $\sigma = \sqrt{n(n + 1)(2n + 1)/24} = 17.61$	419
10.5	Horizontal boxplot of differences of aggression scores	420
10.6	Side-by-side boxplots as well as comparative dotplots for pig weights for diets A and B	428
10.7	Graphical representation of the Wilcoxon rank-sum distribution for $n = m = 10$ superimposed by a normal distribution with $\mu = n(N + 1)/2 = 105$ and $\sigma = \sqrt{nm(N + 1)/12} = 13.22876$	430
10.8	Comparative boxplot for improvements in swim times for high and low-fat diets	433
10.9	Boxplots and density plots of free throw teaching results	437
10.10	Comparative boxplots and density plots for hydrostatic weighing (HWFAT), skin fold measurements (SKFAT), and the Tanita body fat scale (TANFAT)	445
10.11	Histogram of observed goals for <code>Soccer</code> with a superimposed Poisson distribution with $\lambda = 2.5$	451
10.12	Histogram of SAT scores in <code>Grades</code> with superimposed expected values	454
10.13	Graphical illustration of the vertical deviations used to compute the statistic D_n	457
10.14	Graphical illustration of <code>ksdist(n=5, sims=10000, alpha=0.05)</code>	458
10.15	Estimated densities for simple and composite hypotheses from running <code>ksLdist(sims=10000, n=10)</code>	460
10.16	Graphical representation of the bootstrap	471
10.17	Histogram of interarrival times at the M1 motorspeedway	474
10.18	Histogram and quantile-quantile plot of the bootstrapped mean of interarrival times at the M1 motorspeedway	476
10.19	Histogram and quantile-quantile plot of the bootstrapped standard deviation of interarrival times at the M1 motorspeedway	478
10.20	Histogram of the sampling distribution of $\hat{\theta} = \bar{z} - \bar{y}$	482
10.21	Histogram and quantile-quantile plot of $\hat{\theta}^* = \bar{z}^* - \bar{y}^*$ when sampling without replacement (permutation)	483

10.22	Histogram and quantile-quantile plot of $\hat{\theta}^* = \bar{z}^* - \bar{y}^*$ when sampling with replacement (bootstrap)	483
11.1	Representation of a completely randomized design where treatments A, B, and C are assigned at random to six experimental units.	492
11.2	Representation of a randomized complete block design where treatments A, B, and C are assigned at random to three experimental units in each block.	492
11.3	Output from the function <code>oneway.plots(StopDist, tire)</code> including dot-plot, boxplots, and design plot (means) using the data frame <code>Tire</code>	494
11.4	Power for the directional alternative hypothesis $H_1 : \mu_B - \mu_A > 0$ when $\gamma = 2.6$ at the $\alpha = 0.05$ level	504
11.5	Histogram of simulated $F_{3, 20; \lambda=5.25}$ superimposed by the theoretical distribution	506
11.6	Power for detecting treatment differences when $\lambda = 5.25$ at the $\alpha = 0.05$ level	507
11.7	Standardized residuals versus order for <code>mod.aov</code> using the <code>Tire</code> data set	512
11.8	Quantile-quantile plot of the standardized residuals with a superimposed line with a zero intercept and a slope of one for the model <code>mod.aov</code> using the <code>Tire</code> data frame	513
11.9	Plot of the standardized residuals versus the fitted values for <code>mod.aov</code> using the <code>Tire</code> data set	514
11.10	Graphs to assess independence, normality, and constant variance, respectively, created with <code>checking.plots(mod.aov)</code> using the data frame <code>Tire</code>	514
11.11	Transformations in common use with the Box-Cox method: The long dashed line shows data transformed by squaring; the solid, by doing nothing; the dot-dashed, by taking the square root; the dotted, by taking the natural log; and the dashed, by taking the reciprocal	515
11.12	<code>checking.plots()</code> applied to the model <code>FCD.aov (aov(Weight ~ Diet))</code> with the <code>FCD</code> data frame	517
11.13	Box-Cox transformation graph for the model <code>FCD.aov (aov(Weight ~ Diet))</code> with the <code>FCD</code> data frame	518
11.14	<code>checking.plots()</code> applied to the model <code>FCDlog.aov (aov(log(Weight) ~ Diet))</code> using the <code>FCD</code> data frame	518
11.15	These graphs give barplots showing the mean fecundity by line, by contrast 1, and by contrast 2 with individual 95% confidence intervals. The bottom right graph displays the simultaneous 95% confidence intervals for contrast 1 and contrast 2.	528
11.16	Graphical representation of confidence intervals based on Tukey's HSD for the model <code>StopDist ~ tire</code> using the data frame <code>Tire</code>	532
11.17	Multiple comparison boxplot with <code>multcompTs</code> differentiating means based on Tukey's HSD for the model <code>StopDist ~ tire</code> using the data frame <code>Tire</code>	532
11.18	Barplot of mean stopping distance by tire type with superimposed individual 95% confidence intervals for the <code>Tire</code> data frame	534
11.19	Interaction plots of block and treatments using <code>TireWear</code>	542
11.20	The graph resulting from the lattice/Trellis function <code>stripplot()</code> for Example 11.7	542
11.21	Tire wear means due to treatments and blocks	543
11.22	<code>checking.plots()</code> applied to <code>mod.aov</code> from Example 11.7	545
11.23	Simultaneous 95% mean pairwise confidence intervals using Tukey's HSD from Example 11.7	546

11.24	Barplot of the mean wear by tire with superimposed individual 95% confidence intervals from Example 11.7	546
11.25	Graphs resulting from <code>twoway.plots(Microamps, Glass, Phosphor)</code>	551
11.26	Graphs resulting from using <code>checking.plots()</code> on the model <code>mod.TVB</code> from Example 11.8	553
11.27	Interaction plots of glass and phosphor	554
11.28	Tukey HSD 95% family-wise confidence intervals for the model <code>mod.TVB</code>	555
11.29	Barplot of the means for the six treatment combinations of factors <code>Glass</code> and <code>Phosphor</code> with individual superimposed 95% confidence intervals	555
12.1	Graphical representation of simple linear regression model depicting the distribution of Y given x	566
12.2	Scatterplot of <code>gpa</code> versus <code>sat</code> using <code>Grades</code>	575
12.3	Graphical representation of the sum of squares partition	585
12.4	Scatterplots to illustrate values of R^2	586
12.5	Plot of C_p versus p	605
12.6	Residual plots for six different models with different residual patterns	610
12.7	Diagnostic plots for <code>mod1</code> in Figure 12.6	610
12.8	Quantile-quantile plot for <code>mod1</code> in Figure 12.6 on page 610	611
12.9	Residuals versus fitted values for the model <code>HWFAT ~ ABS + TRICEPS</code>	613
12.10	Scatterplot of height (<code>ht</code>) versus weight (<code>wt</code>) for the data set <code>Kinder</code>	616
12.11	Graph of leverage values versus order for regressing height on weight for the data set <code>Kinder</code>	617
12.12	Diagnostic graphs for <code>modk19</code> requested in part (c) of Example 12.17	618
12.13	Diagnostic graphs for <code>modk20</code> requested in part (d) of Example 12.17	619
12.14	Scatterplot of height versus weight for data from <code>Kinder</code> with four superimposed regression lines	620
12.15	Scatterplot, residuals versus fitted values, and quantile-quantile plot of standardized residuals for Y versus x_1 and Y versus $x_1^{0.5}$ models	621
12.16	Scatterplot, residuals versus fitted values, and quantile-quantile plot of standardized residuals for Y versus x_2 and Y versus x_2^2 models	622
12.17	Scatterplot, residuals versus fitted values, and quantile-quantile plot of standardized residuals for Y versus x_3 and Y versus x_3^{-1} models	623
12.18	Box-Cox graph of λ for Example 12.20 on page 626	627
12.19	Scatterplot and residual versus fitted plot of Y_1 versus x_1 ; Box-Cox plot of λ ; scatterplot, residual versus fitted plot, and quantile-quantile plot of $\ln(Y_1)$ versus x_1	628
12.20	Scatterplot and residual versus fitted plot of Y_2 versus x_2 ; Box-Cox plot of λ ; scatterplot, residual versus fitted plot, and quantile-quantile plot of Y_2^{-1} versus x_2	629
12.21	Process of model building with transformations	630
12.22	Small change in x gives a similar small change in $\ln(x)$	631
12.23	Four possible results for a single dummy variable with two levels	634
12.24	Scatterplot of <code>totalprice</code> versus <code>area</code> with the fitted regression line superimposed	636
12.25	Fitted regression lines for Elevators example	638
12.26	Representation of 90% pointwise confidence intervals, 90% prediction intervals, and a 90% confidence band for the regression of <code>gpa</code> on <code>sat</code> using the data in <code>Grades</code>	644
12.27	Joint confidence region for β_2 and β_3 enclosed by the Bonferroni (left graph) and Scheffé (right graph) confidence limits	646

List of Tables

1.1	Body composition (Bodyfat)	10
2.1	Student test scores	41
2.2	Two-way table of Doctor by Ease	50
2.3	Different values for b_0 and b_1 with various regression methods	65
3.1	Probability of two or more students having the same birthday	83
4.1	Comparison of binomial and Poisson probabilities	126
4.2	Standardized scores (data frame Score)	158
5.1	B.S. graduate grades in Linear Algebra and Calculus III	172
5.2	Values used to compute covariance for scatterplots with positive, negative, and zero covariance	182
5.3	Joint probability distribution for X and Y	184
6.1	Finite populations' parameters	203
6.2	Parameters and their corresponding estimators	204
6.3	Finite population parameter estimators and their standard errors	204
6.4	Statistics for samples of size n	205
6.5	Possible samples of size 2 with \bar{x} and s^2 for each sample – random sampling	208
6.6	Sampling distribution of \bar{X} – random sampling	208
6.7	Sampling distribution of S^2 – random sampling	208
6.8	Possible samples of size 2 with \bar{x} and s^2 – simple random sampling	209
6.9	Sampling distribution of \bar{X} – simple random sampling	209
6.10	Sampling distribution of S^2 – simple random sampling	209
6.11	Summary results for sampling without replacement (finite population)	210
6.12	Computed values for random sampling (Case 1) and simple random sampling (Case 2)	210
6.13	Comparison of simulated uniform and exponential distributions to the normal distribution	217
6.14	Output for probability distribution of $(n - 1)S^2/\sigma^2$ example	235
8.1	Weekly spending in dollars (Grocery)	295
8.2	House prices (in thousands of dollars) for three-bedroom/two-bath houses in Watauga County, NC (House)	301
8.3	Apple hardness measurements (Apple)	304
8.4	Mathematical assessment scores for students enrolled in a biostatistics course (Calculus)	307
8.5	Methods for analyzing normal data	313
8.6	Time to complete a complex simulation in minutes (Sundig)	314
9.1	Form of hypothesis tests	341
9.2	Possible outcomes and their consequences for a trial by jury	343

9.3	Relationship between type I and type II errors	344
9.4	Calculation of φ -values for continuous distributions	350
9.5	Duality of $(1 - \alpha) \cdot 100\%$ confidence intervals and α -level tests of significance	352
9.6	Summary for testing the mean when sampling from a normal distribution with known variance (one-sample z -test)	353
9.7	Summary for testing the mean when sampling from a normal distribution with unknown variance (one-sample t -test)	356
9.8	Summary for test for differences in means when taking independent samples from normal distributions with known variances (two-sample z -test)	362
9.9	Summary for test for differences in means when taking independent samples from normal distributions with unknown but assumed equal variances (two-sample pooled t -test)	364
9.10	Summary for test for differences in mean when taking independent samples from normal distributions with unknown and unequal variances (Welch test)	368
9.11	Summary for testing the mean of the differences between two dependent samples when the differences follow a normal distribution with unknown variance (paired t -test)	371
9.12	Summary for testing the population variance when sampling from a normal distribution	374
9.13	Diameters for 20 randomly selected washers (Washer)	374
9.14	Summary for test for equality of variances when sampling from independent normal distributions	376
9.15	Summary for testing the proportion of successes in a binomial experiment (number of successes is $Y \sim Bin(n, \pi)$)	379
9.16	Summary for testing the proportion of successes in a binomial experiment (normal approximation)	384
9.17	Correction factors when $ p - \pi_0 > \frac{1}{2n}$	384
9.18	General form of a 2×2 table	387
9.19	Summary for testing the proportion of successes with Fisher's exact test	388
9.20	Juveniles who failed a vision test classified by delinquency and glasses wearing (Weindling et al., 1986)	388
9.21	Seven possible 2×2 tables that can be constructed where $k = 6$, $m = 9$, and $n = 7$, with their associated probabilities	389
9.22	Observed heart attacks for those physicians taking aspirin and a placebo (Hennekens, 1988)	390
9.23	Summary for testing the differences of the proportions of successes in two binomial experiments (large sample approximation)	393
9.24	Correction factors when $ p_X - p_Y > \frac{1}{2} \left(\frac{1}{m} + \frac{1}{n} \right)$	393
10.1	Summary for testing the median — sign test	405
10.2	Summary for testing the median — approximation to the sign test	406
10.3	Long distance telephone call times in minutes (Phone)	407
10.4	Asymptotic relative efficiency comparisons	410
10.5	Possible sign and rank combinations for the trivial T^+ distribution example	412
10.6	PDF of T^+ for the trivial T^+ distribution example	412
10.7	Summary for testing the median — Wilcoxon signed-rank test	415
10.8	Waiting times in minutes (Wait)	416
10.9	Summary for testing the median — normal approximation to the Wilcoxon signed-rank test	419
10.10	Aggression test scores (Aggression)	421
10.11	Summary for testing equality of medians — Wilcoxon rank-sum test	424

10.12	Summary for testing equality of medians — Mann-Whitney U -test	425
10.13	Summary for testing the difference in two medians — normal approximation to the Wilcoxon rank-sum test	431
10.14	Summary for testing the difference in two medians — normal approximation to the Mann-Whitney U -Test	431
10.15	Sorted improvements in swim times in seconds for high (x) and low (y) fat diets, where rank refers to the rank of the data point in the combined sample of x and y data points (Swimtimes)	433
10.16	Number of successful free throws	437
10.17	Actual free throws with ranks among all free throws	439
10.18	A representation of the ranked data from a randomized complete block design	443
10.19	The first six wrestlers' body fat as measured by the three techniques and their corresponding ranks	445
10.20	Calculating D_n	456
10.21	Twenty-six-year-olds' happiness	463
10.22	Mild dementia treatment results	463
10.23	Contingency table when sampling from a single population	463
10.24	General form and notation used for an $I \times J$ contingency table when sampling from I distinct populations	464
11.1	One-way design	495
11.2	Model, parameters, and estimators for fixed effects, one-way CRD	496
11.3	ANOVA table for one-way completely randomized design	499
11.4	Tire ANOVA table	500
11.5	α_e values for given α_c and various numbers of comparisons	519
11.6	ANOVA table for model Fecundity ~ Line using Drosophila data	524
11.7	ANOVA table for orthogonal contrasts with Drosophila	525
11.8	Shear on frozen carrots by freezer	535
11.9	Frozen carrots ANOVA table	536
11.10	ANOVA table for the randomized complete block design	539
11.11	The tread loss from the TireWear data frame	541
11.12	Sums and estimates for Example 11.7	541
11.13	Tire wear ANOVA table	543
11.14	Layout for observations in a two-factor factorial design	547
11.15	ANOVA table for two-factor factorial design	549
11.16	Data from Hicks (1956) used in Example 11.8	549
11.17	Two-factor factorial design table to complete for (c) of Example 11.8	550
11.18	Two-factor factorial design table COMPLETED for (c) of Example 11.8	552
11.19	ANOVA table for two-factor factorial design for Example 11.8	553
12.1	ANOVA table for simple linear regression	583
12.2	ANOVA table for <code>model.lm <- lm(gpa~sat)</code>	583
12.3	ANOVA table for multiple linear regression	584
12.4	ANOVA table for <code>mod123 <- lm(Y~x1+x2+x3)</code>	588
12.5	ANOVA table for <code>mod321 <- lm(Y~x3+x2+x1)</code>	589
12.6	ANOVA table for Example 12.12 on page 591	592
12.7	Summary of measures of influential observations	615
12.8	Actual change in Jaguar brain weight	632
12.9	Values of \mathbf{X}_{hi} for HSwrestler	644

A.1	Useful Commands When Working with Numeric Vectors	659
A.1	Useful Commands When Working with Numeric Vectors (continued) . . .	660
A.2	S Vector and Matrix Functions	660
A.3	S Functions Used with Arrays, Factors, and Lists	661
A.4	Important Probability Distributions That Work with <i>rdist</i> , <i>pdist</i> , <i>ddist</i> , and <i>qdist</i>	662
A.5	Useful Functions in S for Parametric Inference	662
A.6	Useful Functions in S for Nonparametric Inference	663
A.7	Useful Functions in S for Linear Regression and Analysis of Variance . . .	664
A.8	Useful Contrast Functions in S for Linear Regression and Analysis of Vari- ance	665
A.9	Useful Model Building Functions in S for Linear Regression and Analysis of Variance	665
A.10	Useful Diagnostic Functions in S for Linear Regression and Analysis of Vari- ance	666
A.11	Trellis Functions	666
A.12	Basic Plotting Functions	667
A.13	Graphs Frequently Used with Descriptive Statistics	668
A.14	Commonly Used Graphical Parameters	669

Chapter 1

A Brief Introduction to S

1.1 The Basics of S

S is a system for interactive data analysis that was developed at Bell Laboratories. Two dialects of the S language exist: R, an open source implementation of S available from <http://www.r-project.org>, and S-PLUS, a commercial implementation of S. This book will refer to both R and S-PLUS as simply S. The S language was designed with interactive use in mind. In recent years, the number of new statistical methods and applications that have been developed with this language have caused its dialect R to be considered the “lingua franca” for computational statistics.

R and S-PLUS both run on a number of operating systems. The current text focuses on the use of S for Windows-based machines; however, users of other operating systems should still find the vast majority of the commands valid. Because the Graphical User Interface (GUI) evolves so quickly, the concentration of this text is the command language that has remained fairly static in more recent versions of R and S-PLUS. For basic command-line data analysis, most programs written in R can be translated into S-PLUS, and vice versa. The examples in the text use R (Version 2.6) and S-PLUS (Version 8), the current versions at the time of writing. Code referred to as S will generally work in both R and S-PLUS. If no program is specified, R code is usually present in this text. Comments are often provided to indicate what changes are needed to R code to allow similar commands to run in S-PLUS.

1.2 Using S

When S is launched, the prompt, `>`, is displayed in the **commands window**, indicating that the software is ready to receive input. The convention used in this text is to show what is typed after the command prompt (`>`) followed by the output generated from what is typed. A single expression or assignment is carried out once the user presses the **Enter** key. There is no punctuation required for single expressions and assignments. However, if the user wants to issue multiple expressions and/or assignments on a single line, each expression or assignment must be separated with a semicolon (`;`). To terminate an S session, either type `q()` at the command-line in the commands window or choose **Exit** from the File menu in a GUI environment. On-line help can be accessed by clicking **HELP** in a GUI environment or by typing `help(name of command)` or `?(name of command)` at the command-line. Another way of learning about a function or data set in R is to use the function `example()`. This runs the code in the examples section of the help page. For instance, to execute the code for the function `plot()`, enter the following code at the R prompt:

```
> par(ask=TRUE)
> example(plot)
```

The `par(ask=TRUE)` prompts the user before moving to the next example. The default is `par(ask=FALSE)`; and with that setting, examples available are shown without a pause, making reading code and output nearly impossible. S is a case sensitive language! Consequently, X and x refer to different objects. If the user omits a comma or a parenthesis, or any other type of typographical error occurs when typing at the command-line, a + sign will appear to indicate that the command is incomplete.

1.3 Data Sets

When using S, one should think of data sets as objects. All of the data sets that are created or imported during an S-PLUS session are stored as objects in the `.Data` folder of the projects directory unless they are intentionally erased with the `rm()` command. Data created or imported while using R is stored in memory. The user is prompted at the end of the R session to save the workspace. Consequently, if the computer crashes while R is running, the workspace will be lost. Functions, as well as data sets in S, are considered objects. To obtain a list of objects in the current workspace, type `objects()` or `ls()` at the command-line prompt. The directories S searches when using the functions `objects()` and `ls()` can be displayed by typing `search()` at the command-line prompt. To extract all objects following a particular pattern with R, say all objects starting with E, enter `objects(pattern="^E")`. Likewise, to remove all objects beginning with E, type `remove(list=objects(pattern="^E"))`. To extract all objects following a particular pattern with S-PLUS, say all objects starting with E, enter `objects(pattern="E*")`. Likewise, to remove all objects beginning with E, type `remove(objects(pattern="E*"))`. These last commands are Windows specific. If one enters the same commands on a UNIX system, ALL of the files will be deleted. If, at some point, the entire workspace needs to be cleared, key in `rm(list=ls())`.

Numerous data sets and functions exist in an extensive collection of S packages. An S package is a collection of S functions, data, help files, and other associated files (C, C++, or FORTRAN code) that have been combined into a single entity that can be distributed to other S users. R packages can be downloaded and installed within an R session with the function `install.packages()`. (The Windows version of R has a menu interface to perform this task.) Once a package is installed, it can be loaded with the `library()` function. Data included in a package is immediately available in S by typing the data set name at the command prompt. Contributed R packages can be downloaded and installed from the Comprehensive R Archive Network (CRAN) at <http://www.r-project.org>. A similar type of archive for S-PLUS packages, the Comprehensive S Archival Network (CSAN), is hosted by Insightful at <http://csan.insightful.com>. Consequently, if one wants to use the data set `quine`, which is in the MASS package, first key in

```
> library(MASS)
```

after which one would be able to use the data set `quine`. The data stored in `quine` can be seen by typing `quine` at the command prompt after MASS is loaded. To see all the data sets in a given package, type `data()`. If a more complete description of a particular data file, say `Cars93`, is desired, enter `?Cars93` at the prompt.

The functions and data sets used in this book are available in the PASWR package, which can be downloaded from CRAN at <http://www.r-project.org>. Scripts for each chapter are available from <http://www1.appstate.edu/~arnholta/PASWR> which also contains functions and data sets for using this book with S-PLUS as well as the R PASWR package.

The use of an editor is highly encouraged for viewing and executing the on-line scripts. Tinn-R is a free, Windows-only editor the authors have used extensively that can be found at <http://www.sciviews.org/Tinn-R/>. Using an editor will also help when one is writing and debugging code. For more on editors for a variety of operating systems, see http://www.sciviews.org/_rgui/projects/Editors.html.

1.4 Data Manipulation

1.4.1 S Structures

Before the examples, it will be useful to have a picture in mind of how S structures are related to one another. Figure 1.1 graphically displays the fact that

$$\text{Elements} \subset \text{Vectors} \subset \text{Matrices} \subset \text{Arrays}$$

As the examples progress, it will become clear how S treats these different structures. Broadly speaking, elements are generally numeric, character, or factor. Factors are categorical data whose categories are called levels. For example, “the cities of North Carolina” is a categorical variable. A factor with four levels could be cities with populations between 1 and 1000, 1001 and 10,000, and 10,001 and 100,000, and greater than 100,000 inhabitants.

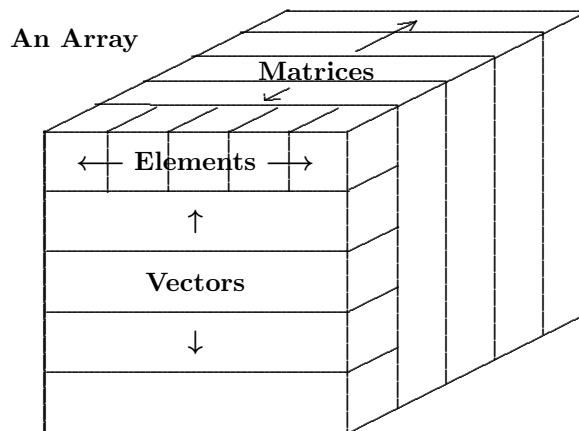


FIGURE 1.1: Structures in S

1.4.2 Mathematical Operations

Arithmetic expressions in S are the usual $+$, $-$, $*$, $/$, and $^$. For instance, to calculate $(7 \times 3) + 12/2 - 7^2 + \sqrt{4}$, enter

```
> (7*3)+12/2-7^2+sqrt(4)
```

and see

```
[1] -20
```

as the output.

Note that the answer to the previous computation is `-20` printed to the right of `[1]`, which indicates the answer starts at the first element of a vector. From this point forward, the command(s) and the output generated will be included in a single section, as both will appear in the commands window, with the understanding that the entire section can be duplicated by entering only what follows the command prompt(s) (`>`). Common functions such as `log()`, `log10()`, `exp()`, `sin()`, `cos()`, `tan()`, and `sqrt()` (square root) are all recognized by S. For a quick reference to commonly used S functions, see Table A.1 on page 659. When working with numeric data, one may want to reduce the number of decimals appearing in the final answer. The function `round(x,2)` rounds the number of decimals to two for the object `x`:

```
> x <- 0.28354
> round(x,2)
[1] 0.28
```

Assigning values to objects in S can be done in several ways. The standard way to assign a value to an object is by using the symbol `<-`. The `=` sign can only be used with R versions 1.6 or later and S-PLUS versions 6 or later. The user should not assign values or objects to reserved letters such as `c`, `q`, `s`, `t`, `C`, `D`, `F`, `I`, or `T`, nor should one write functions with names equal to S functions such as `cor`, `var`, `mean`, or any others. The following commands all assign the value 7 to the object `x`:

```
> x <- 7
> x = 7
```

If `x` had already been assigned a value, the previous value would be overwritten with the new value. Consequently, it is always wise first to ascertain whether an object has an assigned value. To see if an object, say `x`, has an assigned value, type `x` at the command prompt. If the commands window returns

```
> x
Error: Object "x" not found
```

one can assign a value or function to `x` without fear of erasing a preexisting value or function.

1.4.3 Vectors

One special type of object is the **vector**. When working with univariate data, one will often store information in vectors. The S command to create a vector is `c(...)`. To store the values 1.5, 2, and 3 in a vector named `x`, type

```
> x <- c(1.5,2,3)
> x
[1] 1.5 2.0 3.0
```

To square each value in `x`, enter

```
> x^2
[1] 2.25 4.00 9.00
```

To find the position of the entry whose value is 4, enter

```
> which(x^2==4)
[1] 2
```

The S command `c(...)` also works with character data:

```
> y <- c("A", "table", "book")
> y
[1] "A" "table" "book"
```

Some of the more useful commands that are used when working with numeric vectors are included in Table A.1 on page 659.

Two or more vectors can be joined as columns of vectors or rows of vectors. To join two or more column vectors, use the S command `cbind()`. To join two or more row vectors, use the S command `rbind()`. For example, suppose $x = (2, 3, 4, 1)$ and $y = (1, 1, 3, 7)$. If column vectors are desired, use `cbind()`:

```
> x <- c(2,3,4,1)
> y <- c(1,1,3,7)
> cbind(x, y)
      x y
[1,] 2 1
[2,] 3 1
[3,] 4 3
[4,] 1 7
```

If row vectors are desired, use `rbind()`:

```
> rbind(x, y)
  [,1] [,2] [,3] [,4]
x    2    3    4    1
y    1    1    3    7
```

1.4.4 Sequences

The command `seq()` creates a sequence of numbers. Sequences of numbers are often used when creating customized graphs. The three arguments that are typically used with the command `seq()` are the starting value, the ending value, and the incremental value. For example, if a sequence of numbers from 0 to 1 in increments of 0.2 is needed, type `seq(0,1,0.2)`:

```
> seq(0,1,0.2)
[1] 0.0 0.2 0.4 0.6 0.8 1.0
```

When the incremental value is 1, it suffices to use only the starting and ending values of the sequence:

```
> seq(0,8)
[1] 0 1 2 3 4 5 6 7 8
```


An even shorter way to achieve the same result is `0:8`:

```
> 0:8
[1] 0 1 2 3 4 5 6 7 8
```

Decreasing sequences are also possible with commands such as `8:0`:

```
> 8:0
[1] 8 7 6 5 4 3 2 1 0
```

The command `rep(a, n)` is used to repeat the number or character *a*, *n* times. For example,

```
> rep(1,5)
[1] 1 1 1 1 1
```

repeats the value 1 five times. `S` is extremely flexible and allows several commands to be combined:

```
> rep(c(0,"x"), 3)
[1] "0" "x" "0" "x" "0" "x"
> rep(c(1,3,2), length=10)
[1] 1 3 2 1 3 2 1 3 2 1
> c(rep(1,3), rep(2,3), rep(3,3))
[1] 1 1 1 2 2 2 3 3 3
> rep(1:3, rep(3,3))
[1] 1 1 1 2 2 2 3 3 3
```

Specific values in a vector are referenced using square braces `[]`. It is important to keep in mind that `S` uses parentheses `()` with functions and square braces `[]` to reference values in vectors, arrays, and lists. A list is an `S` object whose elements can be of different types (character, numeric, factor, etc.). The following values, stored in `typos`, represent the number of mistakes made per page in the first draft of a research article:

```
> typos <- c(2, 2, 2, 3, 3, 0, 3, 4, 6, 4)
> typos
[1] 2 2 2 3 3 0 3 4 6 4
```

To select the number of mistakes made on the fourth page, type `typos[4]`:

```
> typos[4]
[1] 3
```

To get the number of mistakes made on pages three through six, enter `typos[3:6]`:

```
> typos[3:6]
[1] 2 3 3 0
```

To extract the number of mistakes made on non-continuous pages such as the third, sixth, and tenth pages, key in `typos[c(3,6,10)]`:

```
> typos[c(3,6,10)]
[1] 2 0 4
```

To extract the number of mistakes on every page except the second and third, input `typos[-c(2,3)]`:

```
> typos[-c(2,3)]
[1] 2 3 3 0 3 4 6 4
```

The function `names()` allows the assignment of names to vectors:

```
> x <- c(1,2,3)
> names(x) <- c("A","B","C")
> x
A B C
1 2 3
```

To suppress the names of a vector, type `names(x)<-NULL`:

```
> names(x) <- NULL
> x
[1] 1 2 3
```

1.4.5 Reading Data

S has the ability to read ASCII data stored in external files. S-PLUS can read data stored in a number of other formats, such as MINITABTM worksheets (*.mtw) and/or SPSS files saved as *.sav, while R is slightly more limited with respect to reading other formats. For all but the smallest of data sets, when working with data stored in a format not readable by S, it will almost always prove easier first to save the original data as a text file, and then to read the external file using `read.table()` or `scan()`, although `read.table()` is more user-friendly. For reading data from the console, the function `scan()` may be used.

1.4.5.1 Using scan()

The function `scan()` works well to enter a small amount of data by either typing in the console or using a combination of copying and pasting procedures when the data can be highlighted and copied. To enter the ages for the subjects in Table 1.1 on page 10, one can proceed in two fashions. One can enter all of the ages in one row, or one can enter one age per row. Note that when the values are read into both `age1` and `age2`, the input is terminated by an empty line:

```
> age1 <- scan()
1: 23 23 27 27 39 41 45 49 50 53 53 54 56 57 58 58 60 61
19:
Read 18 items
```

```
> age2 <-scan()
1: 23
2: 23
3: 27
.
.
.
18: 61
19:
Read 18 items
```

1.4.5.2 Using read.table()

The function `read.table()` reads a file in table format (a rectangular data set where the column variables can be quantitative and/or qualitative) and creates a data frame from the external file. When the file contains variable names in the first row, use the argument `header=TRUE`. The default setting in `read.table()` is white space (one or more blank spaces) for field separation. To use other delimiters (commas, periods, etc.) consult the read table help file (`?read.table`). Suppose the data set `Bodyfat` in Table 1.1 on page 10 is a tab delimited ASCII data set stored in a folder named `DATA` under the name `Bodyfat.txt`. To read the data into `S` from the commands window, type

```
> FAT <- read.table("D:/data/Bodyfat.txt", header=TRUE, sep="\t")
> FAT
  age fat sex
1  23  9.5  M
2  23 27.9  F
.   .   .   .
.   .   .   .
.   .   .   .
18 61 34.5  F
```

Note that forward slashes (/) are used to specify the path names. To see the gender for subjects 3 through 6, type

```
> FAT$sex[3:6]
[1] M M F F
Levels: F M
```

For R only The file argument of the `read.table()` command may be a complete url, allowing one to read data into R from the Internet. To read the file `BR.txt` stored on the Internet at

<http://www1.appstate.edu/~arnholta/PASWR/CD/data/Baberuth.txt>,

type

```
> site <-"http://www1.appstate.edu/~arnholta/PASWR/CD/data/Baberuth.txt"
> Baberuth <- read.table(file=url(site), header=TRUE)
> Baberuth[1:5,1:9] # First five rows and nine columns
  Year Team G AB R H X2B X3B HR
1 1914 Bos-A 5 10 1 2 1 0 0
2 1915 Bos-A 42 92 16 29 10 1 4
3 1916 Bos-A 67 136 18 37 5 3 3
4 1917 Bos-A 52 123 14 40 6 3 2
5 1918 Bos-A 95 317 50 95 26 11 11
```

1.4.5.3 Using write()

The function `write()` allows the contents of an `S` data frame or matrix to be saved to an external file in ASCII format. However, one should be aware that information must first be transposed when using the write command. The `S` command to transpose a matrix or data frame is `t(x)`, where `x` is the matrix or data frame of interest. To save the data frame `FAT` to a pen drive, type

```
> write(t(FAT), file="D:/Bodyfat.txt", ncolumns=3)
```

One of the pitfalls to storing information using `write()` is that the file will no longer contain column headings:

```
23  9.5 M
23 27.9 F
27  7.8 M
.   .   .
.   .   .
.   .   .
61 34.5 F
```

The R function `write.table()` avoids many of the inconveniences associated with the S function `write()`. It may be used without transposing the data, it does not lose column headings, and it generally stores the data as a data frame. To save the data frame `FAT` to a pen drive, type `write.table(FAT, file="D:/Bodyfat.txt")` at the R prompt. To read the data stored on the pen drive at a later time, use the function `read.table()`.

1.4.5.4 Using `dump()` and `source()`

Instead of using `write()`, one might use `dump()` to save the contents of an S object, be it a data frame, function, etc. The S function `dump()` takes a vector of names of S objects and produces text representations of the objects in a file. Two of the advantages of using `dump()` are that the dumped file may be read in either R or S-PLUS by using the command `source()` and that the names of the objects are not lost in the writing. A brief example follows that shows how the contents of a vector named `Age` are saved to an external file using R and subsequently opened using the `source()` command in S-PLUS:

```
> dump("Age", file="E:/Age")      # R object Age stored on pen drive.
> source("E:/Age")               # File Age stored on pen drive,
                                # now available in S-PLUS or R
                                # using the same or different
                                # machine.
```

The R function `save()` writes an external representation of R objects to a specified file that can be read on any platform using R. The objects can be read back from the file at a later date by using the function `load()`. If using a point and click interface, the command is labeled `Save Workspace...` and `Load Workspace...`, respectively, found under the `file` drop down menu. S-PLUS allows the user to save data sets in a variety of formats using the `Export Data` command found under the `file` drop down menu.

1.4.6 Logical Operators and Missing Values

The logical operators are `<`, `>`, `<=`, `>=` (less than, greater than, less than or equal to, greater than or equal to), `==` for exact equality, `!=` for exact inequality, `&` for intersection, and `|` for union. The data in Table 1.1 on the following page that are stored in the data frame `Bodyfat` come from a study reported in the *American Journal of Clinical Nutrition* (Mazess et al., 1984) that investigated a new method for measuring body composition.

One way to access variables in a data frame is to use the function `with()`. The structure of `with()` is `with(data frame, expression, ...)`:

Table 1.1: Body composition (Bodyfat)

n	age	% fat	sex	n	age	% fat	sex
1	23	9.5	M	10	53	34.7	F
2	23	27.9	F	11	53	42.0	F
3	27	7.8	M	12	54	29.1	F
4	27	17.8	M	13	56	32.5	F
5	39	31.4	F	14	57	30.3	F
6	41	25.9	F	15	58	33.0	F
7	45	27.4	M	16	58	33.8	F
8	49	25.2	F	17	60	41.1	F
9	50	31.1	F	18	61	34.5	F

```
> with(Bodyfat, fat)
[1] 9.5 27.9 7.8 17.8 31.4 25.9 27.4 25.2 31.1 34.7 42.0 29.1
[13] 32.5 30.3 33.0 33.8 41.1 34.5
```

Suppose one is interested in locating subjects whose fat percentages are less than 25%. This can be accomplished using the `with()` command in conjunction with `fat<25`:

```
> with(Bodyfat, fat < 25)
[1] TRUE FALSE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE
[11] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

To find subjects whose body fat percentages are less than 25% or greater than 35%, enter

```
> with(Bodyfat, fat < 25 | fat > 35)
[1] TRUE FALSE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE
[11] TRUE FALSE FALSE FALSE FALSE FALSE TRUE FALSE
```

To see the fat percentages for subjects with less than 25% fat, type

```
> low.fat <- with(Bodyfat, fat[fat<25])
> low.fat
[1] 9.5 7.8 17.8
```

To remove the subject whose body fat is 7.8 from the previous output, the following may be used:

```
> with(Bodyfat, fat[fat<25 & fat!=7.8])
[1] 9.5 17.8
```

R returns the word `TRUE` or `FALSE` for a logical condition while S-PLUS returns the letters `T` or `F`, where `T` represents true and `F` represents false. From the R output it can be seen that only the first, third, and fourth subjects have fat percentages less than 25%. To select subjects whose fat percentage is less than 25% or greater than 35% without using the `with()` command, attach the data set `Bodyfat` and input `fat<25|fat>35`:

```
> attach(Bodyfat)
> fat<25|fat>35
[1] TRUE FALSE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE
[11] TRUE FALSE FALSE FALSE FALSE FALSE TRUE FALSE
```

Once the data set has been attached, the data set being used remains on the search path until it is detached. If one wants to extract the values from a given vector that satisfy a certain condition, use square braces, `[]`. For example, to store the fat values for all subjects whose fat measured less than 25% in `low.fat`, key in `low.fat <- fat[fat<25]`:

```
> low.fat <- fat[fat<25]
> low.fat
[1] 9.5 7.8 17.8
```

It is also possible to extract values satisfying more complicated logical conditions. For example, to extract all fat percentages that are less than 25% and different from 7.8, enter `fat[fat<25 & fat !=7.8]`:

```
> fat[fat<25 & fat !=7.8]
[1] 9.5 17.8
> detach(Bodyfat)
```

When working with real data, values are often unavailable (the experiment failed, the subject did not show up, the value was lost, etc.). S uses `NA` to denote a missing value or to denote the result of an operation performed on values that contain `NA` values. The function `is.na(x)` returns a logical vector of the same size as `x` that takes on the value `TRUE` if and only if the corresponding element in `x` is `NA`. If `x` is a vector with `NA` values, but only the non-missing values are of interest, the function `!is.na(x)` can be used as shown next:

```
> x <- c(1,6,9,2, NA)
> is.na(x)
[1] FALSE FALSE FALSE FALSE TRUE
> y<-x[!is.na(x)]
> y
[1] 1 6 9 2
```

The following example illustrates how to select the quantitative values of a variable that fulfill a particular character condition (that of receiving treatment A):

```
> x <- c(19,14,15,17,20,23,19,19,21)
> treatment <- c(rep("A",3), rep("B",3), rep("C",3))
> x[treatment=="A"]
[1] 19 14 15
```

To select the value for patients who received `treatment=A` or `treatment=B`, the appropriate command is `x[treatment=="A"|treatment=="B"]`:

```
> x[treatment=="A" | treatment=="B"]
[1] 19 14 15 17 20 23
```

The function `split()` splits the values of a variable `A` according to the categories of a variable `B`:

```
> split(x, treatment)
$A
[1] 19 14 15

$B
[1] 17 20 23

$C
[1] 19 19 21
```

1.4.7 Matrices

Matrices are used to arrange values in rows and columns in a rectangular table. In the following example, different types of barley are in the columns, and different provinces in Spain are in the rows. The entries in the matrix represent the weight in thousands of metric tons for each type of barley produced in a given province. The `barley.data` matrix will be used to illustrate various functions and manipulations that can be applied to a matrix. Given the matrix

$$\begin{pmatrix} 190 & 8 & 22.0 \\ 191 & 4 & 1.7 \\ 223 & 80 & 2.0 \end{pmatrix},$$

the values are written to a matrix (reading across the rows with the command `byrow=TRUE`) with name `barley.data` as follows:

```
> Data <- c(190,8,22,191,4,1.7,223,80,2)
> barley.data <- matrix(Data, nrow=3, byrow=TRUE)
> barley.data
      [,1] [,2] [,3]
[1,] 190   8 22.0
[2,] 191   4  1.7
[3,] 223  80  2.0
```

The matrix's dimensions are computed by typing `dim(barley.data)`:

```
> dim(barley.data)
[1] 3 3
```

The following code creates two objects where the names of the three provinces are assigned to `province`, and the three types of barley to `type`:

```
> province <- c("Navarra", "Zaragoza", "Madrid")
> type <- c("typeA", "typeB", "typeC")
```

Assign the names stored in `province` to the rows of the matrix as follows:

```
> dimnames(barley.data) <- list(province, NULL)
> barley.data
      [,1] [,2] [,3]
Navarra 190   8 22.0
Zaragoza 191   4  1.7
Madrid  223  80  2.0
```

Next, assign the names stored in `type` to the columns of the matrix:

```
> dimnames(barley.data) <- list(NULL, type)
> barley.data
      typeA typeB typeC
[1,]  190   8  22.0
[2,]  191   4   1.7
[3,]  223  80   2.0
```

To assign row and column names simultaneously, the command that should be used is `dimnames(barley.data) <- list(province, type)`:

```
> dimnames(barley.data) <- list(province, type)
> barley.data
      typeA typeB typeC
Navarra 190     8 22.0
Zaragoza 191     4  1.7
Madrid  223    80  2.0
```

One can verify the assigned names with the function `dimnames()`:

```
> dimnames(barley.data)
[[1]]
[1] "Navarra" "Zaragoza" "Madrid"

[[2]]
[1] "typeA" "typeB" "typeC"
```

To delete the row and column name assignments, type

```
> dimnames(barley.data) <- NULL
```

If one is interested in only the second row of data, one can enter

```
> barley.data[2,]
  typeA typeB typeC
   191     4  1.7
```

or

```
> barley.data["Zaragoza", ]
  typeA typeB typeC
   191     4  1.7
```

To see the third column, key in

```
> barley.data[, "typeC"]
Navarra Zaragoza Madrid
     22     1.7     2
```

To add an additional column for a fourth type of barley (`typeD`), use the `cbind()` command:

```
> typeD <- c(2,3.5,2.75)
> barley.data <- cbind(barley.data, typeD)
> rm("typeD")
```

```
> barley.data
      typeA typeB typeC typeD
Navarra 190     8 22.0  2.00
Zaragoza 191     4  1.7  3.50
Madrid  223    80  2.0  2.75
```

(1.1)

The function `apply()` allows the user to apply a function to one or more of the dimensions of an array. To calculate the mean of the columns for the matrix `barley.data`, type `apply(barley.data, 2, mean)`:

```
> apply(barley.data, 2, mean)
      typeA      typeB      typeC      typeD
201.333333 30.666667  8.566667  2.750000
```


The second argument, a 2 in the previous example, tells the function `apply()` to work on the columns. For the function to work on rows, the second argument should be a 1. For example, to find the average barley weight for each province, type `apply(barley.data,1, mean)`:

```
> apply(barley.data, 1, mean)
Navarra Zaragoza Madrid
55.5000 50.0500 76.9375
```

The function `names()` allows the assignment of names to vectors:

```
> x <- c(1,2,3)
> names(x) <- c("A","B","C")
> x
A B C
1 2 3
```

To suppress the names of a vector, type `names(x)<-NULL`:

```
> names(x) <- NULL
> x
[1] 1 2 3
```

1.4.8 Vector and Matrix Operations

Consider the system of equations:

$$\begin{aligned} 3x + 2y + 1z &= 10 \\ 2x - 3y + 1z &= -1 \\ 1x + 1y + 1z &= 6 \end{aligned}$$

This system can be represented with matrices and vectors as

$$\mathbf{Ax} = \mathbf{b}, \text{ where } \mathbf{A} = \begin{bmatrix} 3 & 2 & 1 \\ 2 & -3 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \mathbf{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \text{ and } \mathbf{b} = \begin{bmatrix} 10 \\ -1 \\ 6 \end{bmatrix}$$

To solve this system of equations, enter **A** and **b** into S and type `solve(A, b)` at the command prompt:

```
> A <- matrix(c(3,2,1,2,-3,1,1,1,1), byrow=TRUE, nrow=3)
> A
      [,1] [,2] [,3]
[1,]    3    2    1
[2,]    2   -3    1
[3,]    1    1    1
> b <- matrix(c(10,-1,6), byrow=TRUE, nrow=3)
> b
      [,1]
[1,]   10
[2,]   -1
[3,]    6
> x <- solve(A, b)
```

```
> x
      [,1]
[1,]    1
[2,]    2
[3,]    3
```

The operator `%%` is used for matrix multiplication. If \mathbf{x} is an $(n \times 1)$ column vector, and \mathbf{A} is an $(m \times n)$ matrix, then the product of \mathbf{A} and \mathbf{x} is computed by typing `A%%x`. To verify S's solution, multiply $\mathbf{A} \times \mathbf{x}$, and note that this is equal to \mathbf{b} :

```
> A%%x
      [,1]
[1,]   10
[2,]   -1
[3,]    6
```

Other common functions used with vectors and matrices are included in Table A.2 on page 660.

1.4.9 Arrays

An array generalizes a matrix by extending the number of dimensions to more than two. Consequently, a two-dimensional array of numbers is simply a matrix. If one were to place three (3×3) matrices each in back of the other, the resulting three-dimensional array could be visualized as a cube. Consider a three-dimensional array consisting of 27 elements. Specifically, the elements will be the values 1 through 27. Using the indexing principles illustrated earlier, one can reference an element in the three-dimensional array by specifying the row, column, and depth. For example,

```
> cube <- 1:27
> dim(cube) <- c(3,3,3)
```

assigns the values 1 through 27 into a three-dimensional array. To reference the value in the middle of the cube, one would specify `cube[2,2,2]`:

```
> cube[2,2,2]
[1] 14
```

If any of the indices are left blank, the entire range is reported for that dimension. For example, to extract all the values in the second column with depth 2, type `cube[,2,2]`:

```
> cube[ ,2,2]
[1] 13 14 15
```

Another way to create the array is to specify its elements and dimensions directly. The following code also lists the values in the array so one can see how S processes the entries. Note how `a[, , 1]` can be visualized as the facing matrix in Figure 1.1 on page 3, `a[, , 2]` as the second matrix (slice) in Figure 1.1, and so on:

```
> a <- array(1:27, dim=c(3,3,3))
> a[, , 1]
      [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9
```

```
> a[, , 2]
      [,1] [,2] [,3]
[1,]  10  13  16
[2,]  11  14  17
[3,]  12  15  18
> a[, , 3]
      [,1] [,2] [,3]
[1,]  19  22  25
[2,]  20  23  26
[3,]  21  24  27
```

1.4.10 Lists

A list is an S object whose elements can be of different types (character, numeric, factor, etc.). Lists are used to unite related data that have different structures. For example, a student record might be created by

```
> student <- list(first.name="John", last.name="Smith", major="Biology",
+ semester.hours=15)
```

The object `student` is composed of four components. This can be verified by typing `length(student)` in the commands window. Note that `length()` counts the number of components in a list. Three of the components are character, while the fourth is numeric. The individual components of any list can be extracted by using the `[[` operator or by specifying the name of the list and the name of the component, separated by a dollar sign (`$`). For example, to see the number of semester hours, one might type

```
> student[[4]]
[1] 15
```

or

```
> student$semester.hours
[1] 15
```

Suppose an additional component named `schedule` (a 3×1 array) is added to the list `student`. The second entry in `schedule` can be referenced by typing one of either `student$schedule[2,1]` or `student[[5]][2,1]` since the object `schedule` is at the fifth position in the list `student`.

1.4.11 Data Frames

A data frame is the object most frequently used in S to store data sets. A data frame can handle different types of variables (numeric, factor, logical, etc.) provided they are all the same length. To create a list of variables where the variables are not all of the same type, use the command `data.frame()`. The command `data.frame()` treats the column values in a matrix as the variables and the rows as individual records for each subject in the given variable. Suppose one wishes to code the weather found in the three provinces of the matrix `barley.data`. Unless the user specifies the values in `row.names()`, sequential numbers are assigned to `row.names()` by default. Specifically, one wants to distinguish between provinces that have “continental” weather and those that do not. To add a variable containing character information, use the command `data.frame()` as follows:

```

> cont.weather<-c("no","no","yes")
> city <- data.frame(barley.data, cont.weather)
> rm("cont.weather")
> city
      typeA typeB typeC typeD cont.weather
Navarra  190    8  22.0  2.00          no
Zaragoza 191    4   1.7  3.50          no
Madrid   223   80   2.0  2.75          yes

```

If only barley of typeA is desired, type `city$typeA`:

```

> city$typeA
[1] 190 191 223

```

To make the columns of a data frame available by name, use the command `attach()`. After attaching the data frame `city`, one can view barley of `typeA` by simply typing `typeA`:

```

> attach(city)
> typeA
[1] 190 191 223

```

Note that when finished working with an attached object, one should detach the object using the `detach()` command to avoid inadvertently masking a system object:

```

> detach(city)
> typeC
Error: object "typeC" not found

```

To sort a data frame according to another variable (`typeC` in this example), one can use one of the following: `city[sort.list(city[,3]),]`, `city[order(city[,3]),]`, or `city[order(typeC),]`, all of which produce the same result. Note that `city` will need to be attached again to use the command as given:

```

> attach(city)
> city[sort.list(city[,3]),]
      typeA typeB typeC typeD cont.weather
Zaragoza  191    4   1.7  3.50          no
  Madrid   223   80   2.0  2.75          yes
  Navarra  190    8  22.0  2.00          no
> detach(city)

```

The function `order()` will accept more than one argument to break ties, making it generally more useful than the function `sort.list()`.

1.4.12 Tables

A common use of `table()` is its application to cross-classifying factors to create a table of the counts at each combination of factor levels. In S, factors are simply character vectors. Consider the data set `Cars93`, which contains several numeric and factor variables and is available in the `MASS` package for both R and S-PLUS. To construct a contingency table of Origin by AirBags, use the following S commands:

```

> library(MASS)
> attach(Cars93)

```

```
> table(Origin, AirBags)
      Driver & Passenger Driver only None
USA           9           23  16
non-USA       7           20  18
```

When using three-way contingency tables, `ftable()` provides more compact output than `table()`:

```
> table(Origin, AirBags, DriveTrain)
, , DriveTrain = 4WD
```

```
      AirBags
Origin  Driver & Passenger Driver only None
USA           0           3  2
non-USA       0           2  3
```

```
, , DriveTrain = Front
```

```
      AirBags
Origin  Driver & Passenger Driver only None
USA           6           15  13
non-USA       5           13  15
```

```
, , DriveTrain = Rear
```

```
      AirBags
Origin  Driver & Passenger Driver only None
USA           3           5  1
non-USA       2           5  0
```

```
> ftable(Origin, AirBags, DriveTrain)
      DriveTrain 4WD Front Rear
Origin AirBags
USA     Driver & Passenger      0  6  3
      Driver only              3 15  5
      None                     2 13  1
non-USA Driver & Passenger      0  5  2
      Driver only              2 13  5
      None                     3 15  0
```

Also in R, `margin.table()` and `prop.table()` allow the calculation of totals and proportions by rows or columns:

```
> CT <- table(Origin, AirBags)
> CT
      AirBags
Origin  Driver & Passenger Driver only None
USA           9           23  16
non-USA       7           20  18

> margin.table(CT)      # add all entries in table
[1] 93
```

```

> margin.table(CT,1) # add entries across rows
Origin
  USA non-USA
    48    45
> margin.table(CT,2) # add entries across columns
AirBags
Driver & Passenger      Driver only      None
      16              43              34
> prop.table(CT) # divide each entry by table total
AirBags
Origin  Driver & Passenger Driver only      None
  USA      0.09677419  0.24731183 0.17204301
 non-USA      0.07526882  0.21505376 0.19354839
> prop.table(CT,1) # divide each entry by row total
AirBags
Origin  Driver & Passenger Driver only      None
  USA      0.1875000  0.4791667 0.3333333
 non-USA      0.1555556  0.4444444 0.4000000
> prop.table(CT,2) # divide each entry by column total
AirBags
Origin  Driver & Passenger Driver only      None
  USA      0.5625000  0.5348837 0.4705882
 non-USA      0.4375000  0.4651163 0.5294118

```

1.4.13 Functions Operating on Factors and Lists

In this section, the data set `Cars93` from the `MASS` package is used to illustrate various functions. To find the average `Price` for the vehicles in the `Origin` by `AirBags` table, one might use the function `tapply()` or the function `aggregate()`:

```

> tapply(Price, list(Origin, AirBags), mean)
      Driver & Passenger Driver only      None
USA      24.57778      19.86957 13.33125
non-USA    33.24286      22.78000 13.03333

```

`tapply(x, y, FUN)` applies the function `FUN` to each value in `x` that corresponds to one of the categories in `y`. In this example, `FUN` is the mean. However, in general, `FUN` can be any S or user-defined function. The categories of `y` are the factors created from `list(Origin, AirBags)`, and the `x` is the vector of car prices, `Price`. The final output is a matrix.

The function `aggregate()` is also used to compute the same quantities; however, the output is a data frame:

```

> aggregate(Price, list(Origin, AirBags), mean)
  Group.1      Group.2      x
1  USA Driver & Passenger 24.57778
2 non-USA Driver & Passenger 33.24286
3  USA      Driver only 19.86957
4 non-USA      Driver only 22.78000
5  USA      None 13.33125
6 non-USA      None 13.03333

```

The function `apply(A, MARGIN, FUN)` is used to apply a function *FUN* to the rows or columns of an `array`. For example, given a matrix *A*, the function *FUN* is applied to every row if *MARGIN* = 1 and to every column if *MARGIN* = 2. The function `apply()` is used to compute various statistics with the data frame `Baberuth` as follows:

```
> attach(Baberuth)
> apply(Baberuth[,3:14], 2, mean)
      G      AB      R      H      X2B
113.7727273 381.7727273 98.8181818 130.5909091 23.0000000
      X3B      HR      RBI      SB      BB
 6.1818182 32.4545455 100.5000000  5.5909091 93.7272727
      BA      SLG
0.3228636  0.6340000
```

A summary of the functions covered in this and the previous section can be found in Table A.3 on page 661.

Example 1.1 Assign the values (19, 14, 15, 17, 20, 23, 19, 19, 21, 18) to a vector *x* such that the first five values of *x* are in treatment A and the next five values are in treatment B. Compute the means for the two treatment groups using `tapply()`.

Solution: First assign the values to a vector *x*, where the first five elements are in treatment A and the next five are in treatment B in one of two ways:

```
> x <- c(19,14,15,17,20,23,19,19,21,18)
> treatment <- c(rep("A",5), rep("B",5))
> treatment
[1] "A" "A" "A" "A" "A" "B" "B" "B" "B" "B"
```

or

```
> treatment <- rep(LETTERS[1:2], rep(5,2))
> treatment
[1] "A" "A" "A" "A" "A" "B" "B" "B" "B" "B"
```

Next, use `tapply()` to calculate the means for treatments A and B:

```
> tapply(x, treatment, mean)
 A  B
17 20
```



1.5 Probability Functions

S has four classes of functions that perform probability calculations on all of the distributions covered in this book. These four functions generate random numbers, calculate cumulative probabilities, compute densities, and return quantiles for the specified distributions. Each of the functions has a name beginning with a one-letter code indicating the type of function: *rdist*, *pdist*, *ddist*, and *qdist*, respectively, where *dist* is the S distribution name. Some of the more important probability distributions that work with the functions *rdist*, *pdist*, *ddist*, and *qdist* are listed in Table A.4 on page 662. For example, given vectors *q* and *x* containing quantiles (or percentiles), a vector *p* containing probabilities, and the sample size *n* for a $N(0, 1)$ distribution,

- `pnorm(q, mean=0, sd=1)` computes $\mathbb{P}(X \leq q)$
- `qnorm(p, mean=0, sd=1)` computes x such that $\mathbb{P}(X \leq x) = p$
- `dnorm(x, mean=0, sd=1)` computes $f(x)$
- `rnorm(n, mean=0, sd=1)` returns a random sample of size n from a $N(0, 1)$ distribution.

When illustrating pedagogical concepts, the user will often want to generate the same set of “random” numbers at a later date. To reproduce the same set of “random” numbers, one uses the `set.seed()` function. The `set.seed()` function puts the random number generator in a reproducible state. Verify for yourself that the following code produces identical values stored in the vectors `set1` and `set2`:

```
> set.seed(136)
> set1 <- rbinom(10,10,.3)
> set.seed(136)
> set2 <- rbinom(10,10,.3)
```

This class of functions will also accept a vector as well as a scalar for the function’s arguments. For example, `dpois(x=0:10, lambda=3)`.

1.6 Creating Functions

One of the more attractive features of the S language is the flexibility the user has to modify existing functions and to create new functions. System functions in S are called by typing the name of the function and specifying the arguments being passed to the function inside parentheses. The same principle applies when constructing a new function. The basic structure of a function is

```
> fname <- function(argument1, argument2,...){expression}
```

The **expression** is a mathematical formula that computes its numerical value and/or creates objects based on the user-specified arguments. The result of the expression is computed and subsequently printed in the commands window. When one of the arguments takes a default value in the function definition, there is no need to explicitly type that value when the function is called. The default values for a function can be found in the function’s help file.

Suppose a function to sum the first n natural numbers is needed. The formula to find the sum of the first n natural numbers is $n \times (n + 1)/2$. To create the S function `SUM.N()`, type

```
> SUM.N <- function(n){(n)*(n+1)/2}
```

Using the function `SUM()`, one can see that the sum of the first 10 natural numbers is 55:

```
> SUM.N(10)
[1] 55
```

The function `sum.sq()` sums the squares of the values in a vector or matrix `x`:

```
> sum.sq<-function(x) {sum(x^2)}
```


If one wanted to sum the squared values of each column in the matrix `barley.data` defined in (1.1), one could use

```
> apply(barley.data, 2, sum.sq)
      typeA      typeB      typeC      typeD
122310.0000  6480.0000  490.8900   23.8125
```

1.7 Programming Statements

S, like most programming languages, has the ability to control the execution of code with programming statements such as `for()`, `while()`, `repeat()`, and `break()`. As an example, consider how `for()` is used in the following code to add the values 10, 20, and 30.

```
> sum.a <- 0
> for (i in c(10,20,30)){sum.a <- i + sum.a}
> sum.a
[1] 60
```

In the next section of code, approximate values for converting temperature values from Fahrenheit (60 to 90 by 5 degree increments) to Celsius are given:

```
> for (fahrenheit in seq(60,90,5))
+ print(c(fahrenheit,(fahrenheit-32)*5/9))
[1] 60.00000 15.55556
[1] 65.00000 18.33333
[1] 70.00000 21.11111
[1] 75.00000 23.88889
[1] 80.00000 26.66667
[1] 85.00000 29.44444
[1] 90.00000 32.22222
```

Another way to compute the sum of the first n natural numbers (50 in the code) is to use the function `while()` as follows:

```
> i <- 0; a <- 0; n <- 50
> while (i<n) {i <- i+1; a <- i+a}
> a
[1] 1275
```

When one creates new functions, storing them in a single file can be convenient. By storing all of the functions in a single file, one will be able to read all of them into the S session by typing

```
> source("C:/Sfolder/functions.txt")
```

assuming the functions are all stored in a text file named `functions.txt` in the `Sfolder` of the machine's C drive.

1.8 Graphs

One technique used to summarize numerical data is the proper use of graphs. The S language provides a rich set of commands for creating graphs and altering the default graphical parameters. Tables A.12 on page 667, A.13 on page 668, and A.14 on page 669 outline some of the basic commands used to create graphs and to customize the graphical parameters. In addition to typing commands for graph creation, a large collection of two- and three-dimensional graphs as well as Trellis graphs can be created in S-PLUS from the menu bar by selecting **Graph>2D plot...** or **3D plot...** For further detail on any S function or parameter, the user should seek help from the extensive system help files by typing `help(function.name)`, `?function.name`, `help(par)`, or `?par`.

The S function `plot()` produces an appropriate graph whose form depends on the type of data. The axes, labels, scales, and plotting symbols are all default values chosen automatically, any or all of which may be changed by the user. Changing or adding background color, line types, titles, text, and plotting symbols is all controlled by specifying additional arguments inside S functions such as `plot()` or `hist()`, or by changing certain values in the `par` settings. Table A.14 on page 669 provides a list of some of the more commonly changed graphical parameters. For users who prefer a point and click approach for modifying graphical output, S-PLUS has several buttons on the main menu bar such as Annotation, GraphTools, and Auto Legend.

The following code illustrates the use of various parameters in the S function `plot()` and can be used to recreate Figure 1.2 on the next page. At first the last graph in Figure 1.2 may seem worthless; however, it will often prove useful to create an empty plotting area to which one can later add points, lines, text, and so on. Two of the more frequently used arguments with `par()` are `mfrow` and `mfc01`, which subdivide the plotting region into an array of figure regions. For example, `par(mfrow=c(3,3))` divides the screen into nine figure regions (3 columns by 3 rows). The command `\n` tells R to make a new line in the title.

```
> par(mfrow=c(3,3), pty="m")
> x <- -4:4
> y <- x^2
> plot(x, y, main="Default values with limits \n for x and y axes altered",
+ xlim=c(-8,8), ylim=c(0,20) )
> plot(x, y, pch="x", main="Default plotting character \n changed to x",
+ xlim=c(-8,8), ylim=c(0,20))
> plot(x, y, type="l", main="Lines connecting the data", xlim=c(-8,8),
+ ylim=c(0,20))
> plot(x, y, type="b", main="Both point and lines \n between data",
+ xlim=c(-8,8), ylim=c(0,20))
> plot(x, y, type="h", main="Vertical bars", xlim=c(-8,8), ylim=c(0,20))
> plot(x, y, type="o", main="Overlaid points \n and connected lines",
+ xlim=c(-8,8), ylim=c(0,20))
> plot(x, y, type="s", main="Stairsteps", xlim=c(-8,8), ylim=c(0,20))
> plot(x, y, xlab="X Axis", ylab="Y Axis", main="Basic plot with axes
+ labeled", xlim=c(-8,8), ylim=c(0,20))
> plot(x, y, type="n", main="Empty Graph", xlab="", ylab="", axes=FALSE)
```

The following R code illustrates the use of different plotting symbols, different colors, and different character expansion (`cex`) values and can be used to create a graph similar

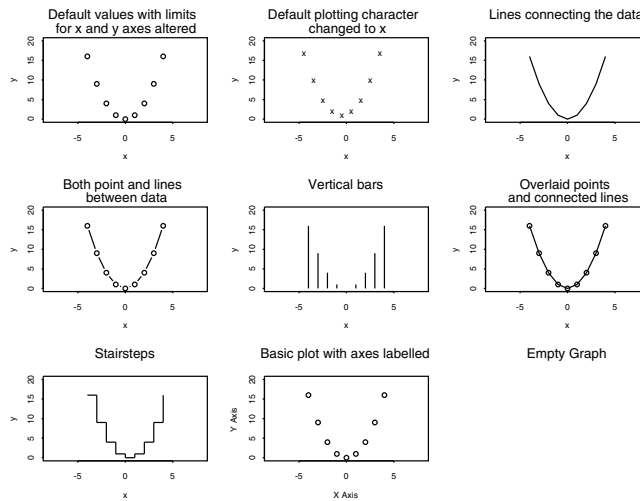


FIGURE 1.2: Examples of the `plot()` function using different values for the parameters `main`, `pch`, `xlim`, `ylim`, `type`, `xlab`, `ylab`, and `axes`.

to Figure 1.3. Color names can be used with a `col=` specification in graphics functions. Numbers or names of colors can be assigned to `col=` as vectors.

```
> plot(1,1, xlim=c(1,16), ylim=c(-1.5,5), type="n", xlab="", ylab="")
> points(seq(1,15,2), rep(4,8), cex=1:8, col=1:8, pch=0:7)
> text(seq(1,15,2), rep(2,8), labels=paste(0:7), cex=1:8, col=1:8)
> points(seq(1,15,2), rep(0,8), pch=(8:15), cex=2)
> text(seq(1,15,2)+.7, rep(0,8), paste(8:15), cex=2)
> points(seq(1,15,2), rep(-1,8), pch=(16:23), cex=2)
> text(seq(1,15,2)+.7, rep(-1,8), paste(16:23), cex=2)
```

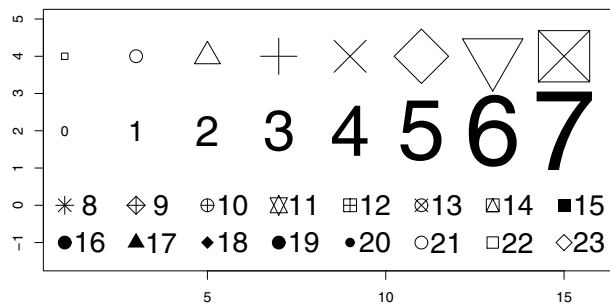


FIGURE 1.3: The numbers in the second row correspond to the plotting symbol directly above them in the first row. The different plotting symbols in the first row and their corresponding numbers in the second row also reflect a character expansion of 1 through 8. The plotting symbols in rows three and four have their corresponding numbers printed to the right.

1.9 Problems

1. Calculate the following numerical results to three decimal places with S:
 - (a) $(7 - 8) + 5^3 - 5 \div 6 + \sqrt{62}$
 - (b) $\ln 3 + \sqrt{2} \sin(\pi) - e^3$
 - (c) $2 \times (5 + 3) - \sqrt{6} + 9^2$
 - (d) $\ln(5) - \exp(2) + 2^3$
 - (e) $(9 \div 2) \times 4 - \sqrt{10} + \ln(6) - \exp(1)$
2. Create a vector named `countby5` that is a sequence of 5 to 100 in steps of 5.
3. Create a vector named `Treatment` with the entries “Treatment One” appearing 20 times, “Treatment Two” appearing 18 times, and “Treatment Three” appearing 22 times.
4. Provide the missing values in `rep(seq(____, _____), _____)` to create the sequence 20, 15, 15, 10, 10, 10, 5, 5, 5, 5.
5. Vectors, sequences, and logical operators
 - (a) Assign the names `x` and `y` to the values 5 and 7, respectively. Find x^y and assign the result to `z`. What is the value stored in `z`?
 - (b) Create the vectors `u = (1, 2, 5, 4)` and `v = (2, 2, 1, 1)` using the `c()` and `scan()` functions.
 - (c) Provide S code to find which component of `u` is equal to 5.
 - (d) Provide S code to give the components of `v` greater than or equal to 2.
 - (e) Find the product `u × v`. How does S perform the operation?
 - (f) Explain what S does when two vectors of unequal length are multiplied together. Specifically, what is `u × c(u, v)`?
 - (g) Provide S code to define a sequence from 1 to 10 called `G` and subsequently to select the first three components of `G`.
 - (h) Use S to define a sequence from 1 to 30 named `J` with an increment of 2 and subsequently to choose the first, third, and eighth values of `J`.
 - (i) Calculate the scalar product (dot product) of $q = (3, 0, 1, 6)$ by $r = (1, 0, 2, 4)$.
 - (j) Define the matrix `X` whose rows are the `u` and `v` vectors from part (b).
 - (k) Define the matrix `Y` whose columns are the `u` and `v` vectors from part (b).
 - (l) Find the matrix product of `X` by `Y` and name it `W`.
 - (m) Provide S code that computes the inverse matrix of `W` and the transpose of that inverse.
6. **Wheat harvested surface in Spain in 2004:** Figure 1.4 on the next page, made with R, depicts the autonomous communities in Spain. The Wheat Table that follows gives the wheat harvested surfaces in 2004 by autonomous communities in Spain measured in hectares. Provide S code to answer all the questions.



FIGURE 1.4: Autonomous communities in Spain

Wheat Table			
community	wheat.surface	community	wheat.surface
Galicia	18817	Castilla y León	619858
Asturias	65	Madrid	13118
Cantabria	440	Castilla-La Mancha	263424
País Vasco	25143	C. Valenciana	6111
Navarra	66326	Región de Murcia	9500
La Rioja	34214	Extremadura	143250
Aragón	311479	Andalucía	558292
Cataluña	74206	Islas Canarias	100
Islas Baleares	7203		

- Create the variables `community` and `wheat.surface` from the Wheat Table in this problem. Store both variables in a `data.frame` named `wheatSpain`.
- Find the maximum, the minimum, and the range for the variable `wheat.surface`.
- Which community has the largest harvested wheat surface?
- Sort the autonomous communities by harvested surface in ascending order.
- Sort the autonomous communities by harvested surfaces in descending order.
- Create a new file called `wheat.c` where `Asturias` has been removed.
- Add `Asturias` back to the file `wheat.c`.
- Create in `wheat.c` a new variable called `acre` indicating the harvested surface in acres (1 acre = 0.40468564224 hectares).
- What is the total harvested surface in hectares and in acres in Spain in 2004?
- Define in `wheat.c` the `row.names()` using the names of the communities. Remove the `community` variable from `wheat.c`.
- What percent of the autonomous communities have a harvested wheat surface greater than the mean wheat surface area?

- (l) Sort `wheat.c` by autonomous communities' names (`row.names()`).
 - (m) Determine the communities with less than 40,000 acres of harvested surface and find their total harvested surface in hectares and acres.
 - (n) Create a new file called `wheat.sum` where the autonomous communities that have less than 40,000 acres of harvested surface have their actual names replaced by "less than 40,000."
 - (o) Use the function `dump()` on `wheat.c`, storing the results in a new file named `wheat.txt`. Remove `wheat.c` from your path and check that you can recover it from `wheat.txt`.
 - (p) Create a text file called `wheat.dat` from the `wheat.sum` file using the command `write.table()`. Explain the differences between `wheat.txt` and `wheat.dat`.
 - (q) Use the command `read.table()` to read the file `wheat.dat`.
7. The data frame `wheatUSA2004` from the `PASWR` package has the USA wheat harvested crop surfaces in 2004 by states. It has two variables, `STATE` for the state and `ACRES` for thousands of acres.
- (a) Attach the data frame `wheatUSA2004` and use the function `row.names()` to define the states as the row names.
 - (b) Define a new variable called `ha` for the surface area given in hectares where 1 acre = 0.40468564224 hectares.
 - (c) Sort the file according to the harvested surface area in acres.
 - (d) Which states fall in the top 10% of states for harvested surface area?
 - (e) Save the contents of `wheatUSA2004` in a new file called `wheatUSA.txt` in your favorite directory. Then, remove `wheatUSA2004` from your workspace, and check that the contents of `wheatUSA2004` can be recovered from `wheatUSA.txt`.
 - (f) Use the command `write.table()` to store the contents of `wheatUSA2004` in a file with the name `wheatUSA.dat`. Explain the differences between storing `wheatUSA2004` using `dump()` and using `write.table()`.
 - (g) Find the total harvested surface area in acres for the bottom 10% of the states.
8. Use the data frame `vit2005` in the `PASWR` package, which contains data on the 218 used flats sold in Vitoria (Spain) in 2005 to answer the following questions. A description of the variables can be obtained from the help file for this data frame.
- (a) Create a table of the number of flats according to the number of garages.
 - (b) Find the mean of `totalprice` according to the number of garages.
 - (c) Create a frequency table of flats using the categories: number of garages and number of elevators.
 - (d) Find the mean flat price (total price) for each of the cells of the table created in part (c).
 - (e) What command will select only the flats having at least one garage?
 - (f) Define a new file called `data.c` with the flats that have `category="3B"` and have an elevator.
 - (g) Find the mean of `totalprice` and the mean of `area` using the information in `data.c`.

Source: Departamento de Economía y Hacienda de la Diputación Foral de Álava and LKS Tasaciones, 2005.

9. Use the data frame `EPIDURALf` to answer the following questions:
 - (a) How many patients have been treated with the `Hamstring Stretch`?
 - (b) What proportion of the patients treated with `Hamstring Stretch` were classified as each of `Easy`, `Difficult`, and `Impossible`?
 - (c) What proportion of the patients classified as `Easy` to palpate were assigned to the `Traditional Sitting` position?
 - (d) What is the mean weight for each cell in a contingency table created with the variables `Ease` and `Treatment`?
 - (e) What proportion of the patients have a body mass index ($\text{BMI} = \text{kg}/(\text{cm}/100)^2$) less than 25 and are classified as `Easy` to palpate?
10. The millions of tourists visiting Spain in 2003, 2004, and 2005 according their nationalities are given in the following table:

Nationality	2003	2004	2005
Germany	9.303	9.536	9.918
France	7.959	7.736	8.875
Great Britain	15.224	15.629	16.090
USA	0.905	0.894	0.883
Rest of the world	17.463	18.635	20.148

- (a) Store the values in this table in a matrix with the name `tourists`.
 - (b) Calculate the totals of the rows.
 - (c) Calculate the totals of the columns.
11. Use a `for` loop to convert a sequence of temperatures (18 to 28 by 2) from degrees centigrade to degrees Fahrenheit.
12. If 1 km = 0.6214 miles, 1 hectare = 2.471 acres, and 1 L = 0.22 gallons, write a function that converts kilometers, hectares, and liters into miles, acres, and gallons, respectively. Use the function to convert 10.2 km, 22.4 hectares, and 13.5 L.

Chapter 2

Exploring Data

2.1 What Is Statistics?

You may be wondering “What is statistics?”, “Who uses it?”, and “Why do I need to study this material?” Statistics is the process of finding out more about a topic by collecting information and then trying to make sense out of that information. In essence, statistics is concerned with methods for collecting, organizing, summarizing, presenting, and analyzing data. Data laden information is present in virtually every sector of society, and the need to make sense out of our surroundings is a basic human need. More to the point of why you, the reader, might need to study this material can be answered in one of two ways. First, you are required to study this material as part of your major because there are certain topics that are deemed important by your teachers. Second, you desire to have some modicum of control in decision making and want to learn more about how probability and statistics help people, corporations, and governmental agencies make decisions/policies. Even if your reason for reading this material is because it is required, it is a fervent hope that your ability to make sound decisions is strengthened through the material in this book.

2.2 Data

Data, according to *The American Heritage Dictionary*, are “Information, especially information organized for analysis or used as the basis for a decision.” A characteristic that is being studied in a statistical problem is called a **variable**. A variable will be either **qualitative** or **quantitative**. When a variable is qualitative, it is essentially defining groups or categories. When the categories have no ordering the variable is called **nominal**. For example, the variable gender can take on the values male and female or the variable “music preference” could have values such as “classical,” “jazz,” “rock,” or “other.” When the categories have a distinct ordering, the variable is called **ordinal**. Such a variable might be educational level with values elementary school, high school, college graduate, graduate or professional school. Values on a scale can be either interval or ratio. Interval data have interpretable distances, while ratio data have a true zero. A variable that is quantitative (numeric) may be either **discrete** or **continuous**. A discrete variable is a numerical variable that can assume a finite number or at most a countably infinite number of values. Such variables include the number of people arriving at a bank on Thursday, students in a class, or dogs in the pound. A continuous variable is a numerical variable that can assume an infinite number of values associated with the numbers on an interval of the real number line, for example, the height of a tree, the life of a light bulb, the weight of an apple. An important distinction between discrete and continuous variables is that discrete variables

can take on the same value repeatedly while continuous variables have few or no repeated values. It is important to be able to distinguish between different types of variables since methods for viewing and summarizing data are dependent on variable type. More to the point, it will be imperative to distinguish between qualitative (categorical) variables and quantitative (numerical) variables.

When a data set consists of a single variable, it is called a **univariate** data set. When there are two variables in the data set, the data set is called a **bivariate** data set; and when there are two or more variables, the data set is called a **multivariate** data set. In the remainder of this section, the discussion will cover univariate variables. Recall that a qualitative variable defines categories or groups. The membership in these categories is summarized with tables and graphically illustrated with bar graphs.

2.3 Displaying Qualitative Data

2.3.1 Tables

A table that lists the different groups of categorical data and the corresponding frequencies with which they occur is called a **frequency table**. Qualitative information is typically presented in the form of a frequency table. The S function `table()` can be used to create various types of tables.

Example 2.1 Suppose the letter grades of an English essay in a small class are A, D, C, D, C, C, C, C, F, and B. Create both a frequency table showing the numbers and a relative frequency table showing the proportions of the various grades.

Solution: First, the character data are read into a vector named `Grades`. Then, the S function `table()` is applied to `Grades`:

```
> Grades <- c("A", "D", "C", "D", "C", "C", "C", "C", "F", "B")
> Grades
[1] "A" "D" "C" "D" "C" "C" "C" "C" "F" "B"
> table(Grades)
Grades
A B C D F
1 1 5 2 1
> table(Grades)/10      # Relative frequency table
Grades
  A   B   C   D   F
0.1 0.1 0.5 0.2 0.1
```

Clearly, there is no need for a computer with such a small data set; however, tables for much larger data sets can be created with no more work than that required for this small data set. ■

Example 2.2 The `quine` data frame in the `MASS` package has information on children from Walgett, New South Wales, Australia, that were classified by `Culture`, `Age`, `Sex`, and `Learner` status including the number of `Days` absent from school in a particular school year. Use the function `table()` to create a frequency table for the variable `Age`.

Solution: To gain access to information stored in `MASS`, first load the package and attach the data frame `quine`:

```

> library(MASS)
> attach(quine)
> table(Age)
Age
F0 F1 F2 F3
27 46 40 33

```



2.3.2 Barplots

One of the better graphical methods to summarize categorical data is with a **barplot**. Barplots are also known as bar charts or bar graphs. The S function `barplot()` is used to create barplots using a summarized version of the data, often the result of the `table()` function. This summarized form of the data can be either frequencies or percentages. Regardless of whether one uses frequencies or percentages, the resulting shape looks identical, but the scales on the y -axes are different.

Example 2.3 Construct barplots for the variables `Grades` used in Example 2.1 and `Age` in the `quine` data set from the `MASS` package in Example 2.2 on the facing page using both frequencies and proportions.

Solution: Before creating any barplots, the device region is split into four smaller regions with the command `par(mfrow=c(2,2))`:

```

> par(mfrow=c(2,2))
> barplot(table(Grades), col=3, xlab="Grades", ylab="Frequency")
> barplot(table(Grades)/length(Grades), col=3, xlab="Grades", ylab=
+ "Proportion")
> barplot(table(Age), col=7, xlab="Age", ylab="Frequency")
> barplot(table(Age)/length(Age), col=7, xlab="Age", ylab="Proportion")

```

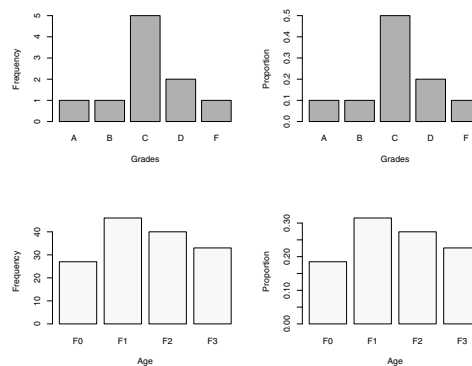


FIGURE 2.1: Graphical representation of the data in `Grades` and `Age` with the function `barplot()`



2.3.3 Dot Charts

An equally effective way to display qualitative data is by using a **dot chart**. Dot charts are also called Cleveland dotplots. A dot chart shows the values of the variables of interest (levels of the qualitative variable) as dots in a horizontal display over the range of the data. The S command to create a dot chart is `dotchart(data)`, where `data` is a vector containing frequencies for all the different levels of a variable. When working with un-summarized data, one way to prepare the data for a `dotchart()` is first to summarize the data with the command `table()`. The optional arguments for `dotchart()` in R and S-PLUS are different, and the user should consult the respective documentation for further assistance.

Example 2.4 Construct dot charts for the variables `Grades` from Example 2.1 and `Age` used in the `quine` data set from the `MASS` package in Example 2.2 on page 30.

Solution: Before creating any dot charts, the device region is split into two smaller regions with the command `par(mfrow=c(2,1))`:

```
> par(mfrow=c(1,2))
> dotchart(table(Grades))
> dotchart(table(Age))
```

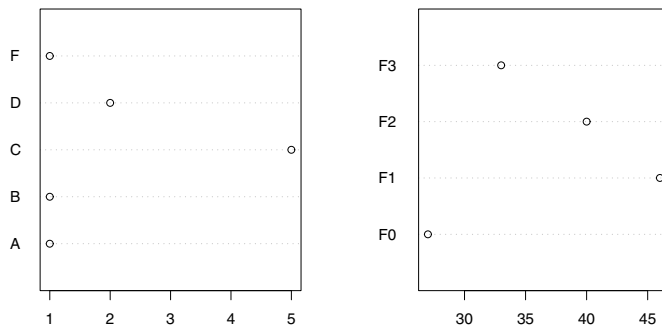


FIGURE 2.2: Graphical representation of the data in `Grades` and `Age` with the function `dotchart()` ■

2.3.4 Pie Charts

Pie charts represent the relative frequencies or percentages of the levels of a categorical variable with wedges of a pie (circle). While the media often use **pie charts** to display qualitative data, the pie chart has fallen out of favor with most statisticians. Pie charts are most useful when the emphasis is on each category in relation to the total. When such an emphasis is not the primary point of the graphic, a bar chart or a dot chart should be used.

Example 2.5 Construct pie charts for the variables `Grades` in Example 2.1 and `Age` from the `quine` data set in the `MASS` package used in Example 2.2 on page 30.

Solution: Before creating any pie charts, the device region is split into two regions with the command `par(mfrow=c(2,1))`:

```

> par(mfrow=c(1,2))
> pie(table(Grades))
> title("Grades")
> pie(table(Age))
> title("Age")

```

The graph depicted in Figure 2.3 was produced in R with the additional arguments `radius=2.5` and `col=gray(c(.1,.4,.7,.8,.95))`.

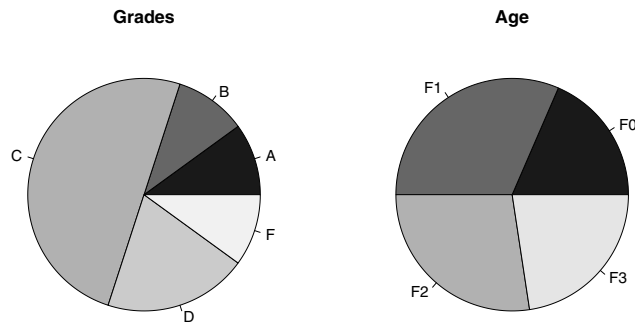


FIGURE 2.3: Graphical representation of the data in `Grades` and `Age` with the function `pie()` ■

2.4 Displaying Quantitative Data

When presented with quantitative data, knowing three facts about the data, namely, its shape, center, and spread, will be a great start in making some sense of the numbers. Some of the more common distribution shapes are shown in Figure 2.4 on the following page. Of the nine different shapes in Figure 2.4, all are symmetric with the exception of the second and the eighth graphs, which are characterized as skewed to the right and skewed to the left, respectively. Of the nine different shapes in Figure 2.4, all are unimodal with the exception of the first, the fourth, and the ninth graphs, which are characterized as bimodal, uniform, and multi-modal, respectively. One final highlight: When presented with a symmetric unimodal data set, it will be important to classify the distribution as either short-tailed, long-tailed, or normal. The fourth and the sixth graphs, in addition to being symmetric, are also short-tailed. What follows are graphical tools that can help in assessing the shape, center, and spread of a data set. As a general rule, the shape of the data dictates the most appropriate measures of center and spread for that data set.

2.4.1 Stem-and-Leaf Plots

One way to get a quick impression of the data is to use a **stem-and-leaf plot**. When a stem-and-leaf plot is constructed, each observation is split into a stem and a leaf. Regardless of where the observation is split, the leaf in a stem-and-leaf plot is represented with a single

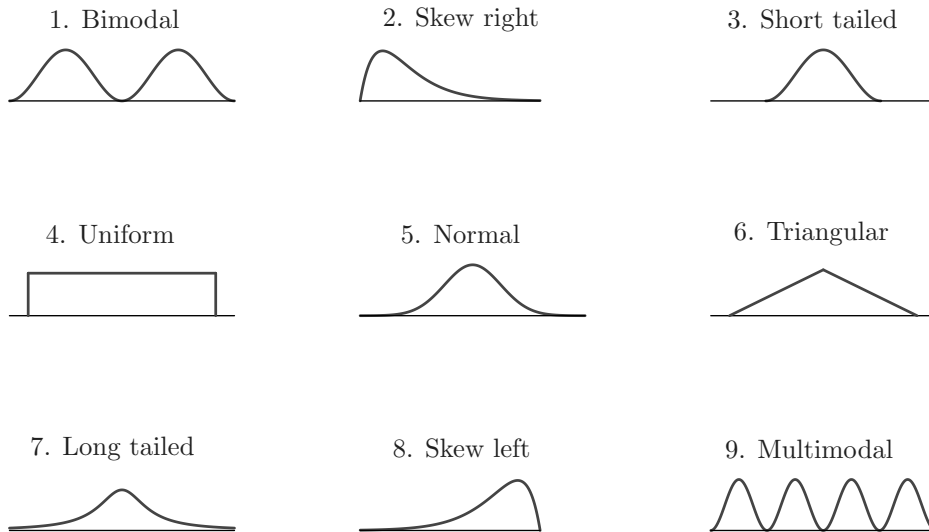


FIGURE 2.4: Nine different graphs labeled according to their shape

digit. Although it is possible to use a stem-and-leaf plot with a moderately sized data set (more than 100 values), the plot becomes increasingly hard to read as the number of values plotted increases. Consequently, it is recommended that stem-and-leaf plots be used graphically to illustrate smallish data sets (less than 100 values). The `S` command to create a stem-and-leaf plot is `stem(x)`, where `x` is a numeric vector.

Example 2.6 Use the data frame `Baberuth` to construct a stem-and-leaf plot for the number of home runs (HR) Babe Ruth hit while he played for the New York Yankees.

Solution: A quick glance at the data frame `Baberuth` shows that Babe Ruth played for the New York Yankees for his seventh through twenty-first seasons. The information in `HR` is for Babe Ruth's entire (22 seasons) professional career. To extract the home runs he hit while he was a New York Yankee, use `HR[Team=="NY-A"]` or `HR[7:21]` (seventh through twenty-first season home runs):

```
> attach(Baberuth)                # Assumes package PASWR is loaded
> NYYHR <- HR[Team=="NY-A"]
> NYYHR
[1] 54 59 35 41 46 25 47 60 54 46 49 46 41 34 22
> stem(NYYHR)
```

The decimal point is 1 digit(s) to the right of the |

```
2 | 25
3 | 45
4 | 1166679
5 | 449
6 | 0
```

```
> detach(Baberuth)
```

In this example, see how the stems 2–6 represent the values twenty through sixty and the leaves represent the second digit of the numbers in HR. Reading the first row of the stem-and-leaf plot notice the values 22 and 25. The stem-and-leaf plot reveals a fairly symmetric distribution. ■

2.4.2 Strip Charts (R Only)

An alternative to the stem-and-leaf plot is a **strip chart** (also referred to as a dotplot by many authors). A strip chart plots values along a line. The R function `stripchart()` will stack the tied observations in a column at each value observed along a line that covers the range of the data when given the argument `method="stack"`. The function requires the data to be a vector, a list of vectors, or a formula of the form `x~g`, where values are in a vector `x` and groups are in a vector `g`. Strip charts are often useful for comparing the distribution of a quantitative variable at different qualitative levels (groups).

Example 2.7 Use the data frame `Baberuth` to

- Construct a strip chart of the number of home runs Babe Ruth hit while playing for the New York Yankees.
- Create a strip chart of the number of home runs Babe Ruth hit per season according to the team for which he was playing. Based on the strip chart, when Babe Ruth played, for which team did he generally hit more home runs per season?

Solution: (a) Figure 2.5 on the next page is a strip chart of the number of home runs Babe Ruth hit while playing for the New York Yankees. The code to construct this graph is

```
> attach(Baberuth)
> Baberuth[1:5,]      # equivalently heads(Baberuth, n=5)
  Year Team  G  AB  R  H X2B X3B HR RBI SB BB  BA  SLG
1 1914 Bos-A  5  10  1  2   1  0  0  0  0  0 0.200 0.300
2 1915 Bos-A 42  92 16 29  10  1  4 21  0  9 0.315 0.576
3 1916 Bos-A 67 136 18 37   5  3  3 16  0 10 0.272 0.419
4 1917 Bos-A 52 123 14 40   6  3  2 12  0 12 0.325 0.472
5 1918 Bos-A 95 317 50 95  26 11 11 66  6 58 0.300 0.555
> NYHR <- HR[7:21]    # Extracts the 7th through 21st season HR values.
> stripchart(NYHR, xlab="Home runs per season", pch=1, method="stack",
+ main="Dotplot of home runs while a New York Yankee")
```

(b) Figure 2.6 on the following page is a strip chart of the number of home runs Babe Ruth hit per season according to the team for which he was playing. The code to construct this graph is

```
> par(mfrow=c(1,2), pty="s")
> stripchart(HR~Team, pch=1, method="stack",
+ main="Dotplot of home runs \n by team",
+ xlab="Home runs per season")
> par(las=1) # Makes labels horizontal
> stripchart(HR~Team, pch=19, col=c("red","green","blue"),
+ method="stack", main="Color dotplot of home runs \n by team",
+ xlab="Home runs per season")
> par(mfrow=c(1,1), las=0, pty="m")
> detach(Baberuth)
```

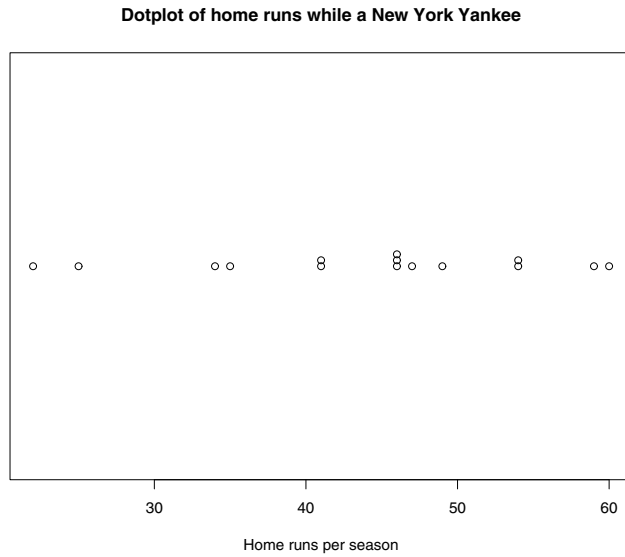


FIGURE 2.5: Strip chart of the number of home runs Babe Ruth hit while playing for the New York Yankees

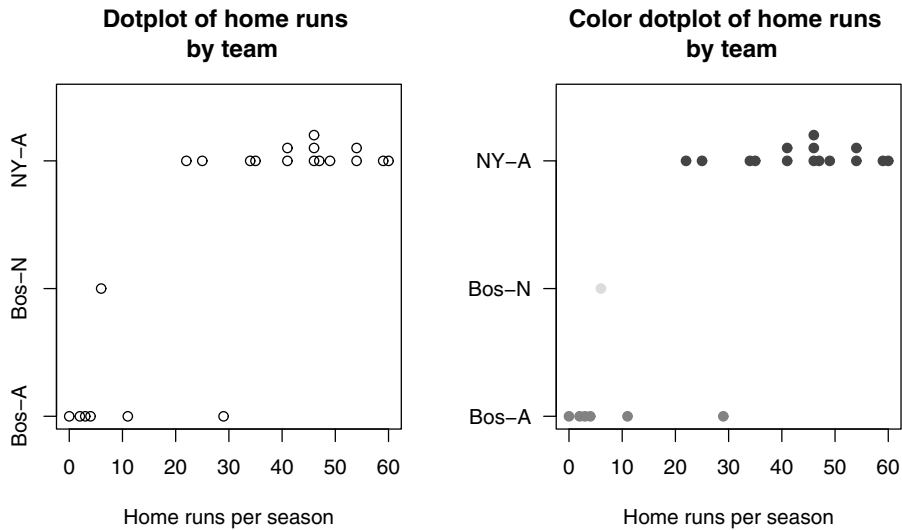


FIGURE 2.6: Strip chart of the number of home runs Babe Ruth hit per season according to the team for which he was playing

2.4.3 Histograms

The **histogram** is a graphical means of illustrating quantitative (numerical) data. Although the barplot and the histogram look similar, the barplot is used for qualitative data while

the histogram is used for numerical data. Yet, the bins that either the user specifies or those that `S` uses by default are in essence categories. Histograms created in `S` with the function `hist(x)`, where `x` is a numeric vector, are by default frequency histograms. To create density histograms, use the optional argument `prob=TRUE`. A density histogram has a total area of one.

Example 2.8 Construct a histogram that resembles the stem-and-leaf plot from Example 2.6 using the `Baberuth` data.

Solution: The first histogram uses the default arguments for `hist()`. Since the bins `S` uses are of the form `()`, the default histogram does not resemble the stem-and-leaf plot. To change the bins to the form `[]` in R, use the argument `right=FALSE`:

```
> attach(Baberuth)
> par(mfrow=c(1,2))
> bin <- seq(20,70,10) # Creating bins 20-70 by 10
> hist(HR[7:21], breaks=bin, xlab="Home Runs")
> hist(HR[7:21], breaks=bin, right=FALSE, xlab="Home Runs") # R
> detach(Baberuth)
```

The graph depicted in Figure 2.7 was produced in R with commands similar to those given. One way to produce the second graph in `S-PLUS` is to use a slight fudge factor when creating the bins, such as `bin <- seq(20,70,10)-0.00001`.

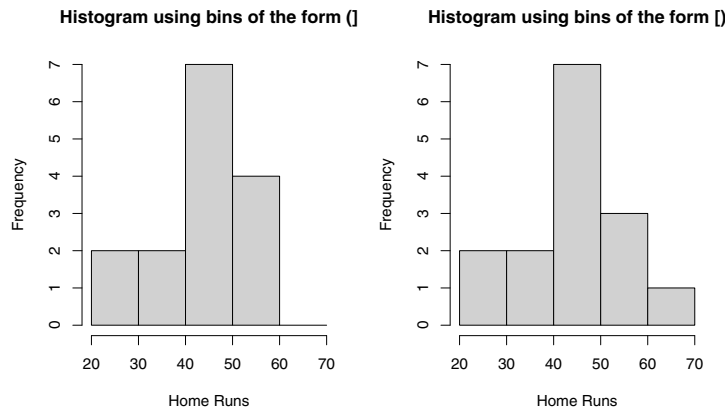


FIGURE 2.7: Histograms created using different bin definitions for the number of home runs hit by Babe Ruth while playing for the New York Yankees ■

One of the problems with using histograms to describe the shape of the data is the arbitrary nature of the bin width. In Example 2.8, it was seen how simply including or excluding an end point changed the histogram. Consider the differences among the shapes of the histograms in Figure 2.8 on the next page produced by simply altering the bin width. The data set used to produce Figure 2.8 on the following page is `geyser`, available in the `MASS` package. A much better choice to get an idea of what the shape of a distribution looks like is to use a **density estimate**. The `S` function `density(x)`, where `x` is a numeric

vector, can be used to create a density estimate. Basically, a density estimate uses shapes with $\frac{1}{n}$ area added up at each point in the data set to create a graph with area 1. The resulting shape is a density estimate. The result of the density estimate can be viewed with either the `plot()` or `lines()` function. Recall that `plot()` is a high-level function while `lines()` is a low-level function. That is, `plot()` will create a graph while `lines()` will add to an existing graph.

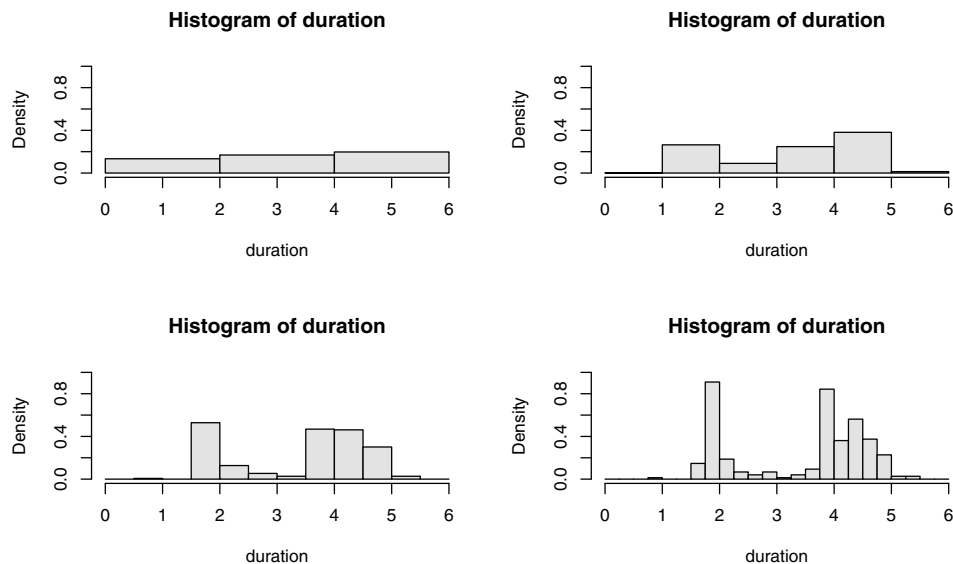


FIGURE 2.8: Histograms created using different bin definitions for the eruption duration of Old Faithful

Example 2.9 Construct a density histogram of the waiting time until the next eruption using the data frame `geyser` available in the `MASS` package. Superimpose a density estimate over the density histogram. In the same graph, show the estimated density without showing the histogram.

Solution: Note that to superimpose a density over a histogram, the histogram must be a density histogram. Recall that density histograms are produced with the optional argument `prob=TRUE`:

```
> library(MASS)
> par(mfrow=c(1,2))      # Make device region 1 by 2
> attach(geyser)
> hist(waiting, prob=TRUE)
> lines(density(waiting)) # Add density to Histogram
> plot(density(waiting)) # Create density by itself
> detach(geyser)
```

Based on the density estimates, it appears there are two modes for waiting time until the next eruption. It seems one will usually have to wait close to either 50 or 80 minutes until the next eruption. ■

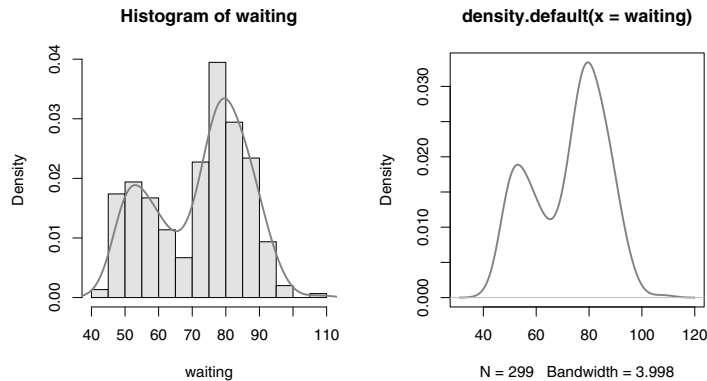


FIGURE 2.9: Histogram of waiting time between Old Faithful eruptions with superimposed density estimate as well as a density plot

2.5 Summary Measures of Location

One of the main objectives of statistics is to make inference to a population based on information obtained from a sample. Since it can be overwhelming to work with the entire population and/or sample, summary measures are introduced to help characterize the data at hand. These summary measures may apply to either the population or to the sample. Numerical summaries of the population are called **parameters** while numerical summaries of the sample are called **statistics**. More formal definitions of both parameters and statistics will be given later. Measures of central location are introduced first. The measures covered are generally familiar to the reader from everyday usage. Specifically, the **mean**, the **trimmed mean**, and the **median** are introduced. Other measures of location addressed include quartiles, hinges, and quantiles.

2.5.1 The Mean

The most common measure of center is the average, which locates the balance point of the distribution or data. The mean is an appropriate measure of center for symmetric distributions; however, it is not appropriate for skewed distributions. In statistics, the average of a sample is called the **sample mean** and is denoted by \bar{x} . Given some numeric data x_1, x_2, \dots, x_n , the sample mean is defined as

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \sum_{i=1}^n \frac{x_i}{n} \quad (2.1)$$

The S function `mean(x)` will compute the mean of a data vector `x`. Additional arguments to `mean(x)` include `na.rm=TRUE`, for removal of missing values, and `trim=`, to compute a trimmed mean. The trimmed mean is generally used to estimate the center when working with long-tailed distributions. When a $p\%$ trimmed mean is computed, $p\%$ of the sorted data is deleted from each end of the distribution, and a mean is computed from the remaining

values. When $p \times n$ is not an integer, the integer portion, $(\lfloor p \times n \rfloor)$, should be deleted from each end of the sorted values and the mean computed from the remaining values.

Example 2.10 Compute the mean number of home runs per season Babe Ruth hit while playing for the New York Yankees. Compute a 5%, a 10%, a 15%, and a 50% trimmed mean for the number of home runs per season Babe Ruth hit while playing for the New York Yankees using the information stored in the data frame **Baberuth**.

Solution: In Example 2.6 on page 34, the variable **NYHR** was created that contained the number of home runs Babe Ruth made while playing for the New York Yankees. If **NYHR** is no longer available, recreate it with the command `NYHR <- HR[7:21]` once the data frame **Baberuth** has been attached. Since there are 15 values in **NYHR**, to compute 5%, 10%, 15%, and 50% trimmed means, $\lfloor 0.05 \times 15 \rfloor = \lfloor 0.75 \rfloor = 0$, $\lfloor 0.10 \times 15 \rfloor = \lfloor 1.5 \rfloor = 1$, $\lfloor 0.15 \times 15 \rfloor = \lfloor 2.25 \rfloor = 2$, and $\lfloor 0.50 \times 15 \rfloor = \lfloor 7.5 \rfloor = 7$ values, respectively, will need to be deleted from the sorted values of **NYHR** before computing means on the remaining values. A second solution is also presented using the S function `mean()` using the `trim=` argument:

```
> attach(Baberuth)
> NYHR <- HR[7:21]
> NYHR
[1] 54 59 35 41 46 25 47 60 54 46 49 46 41 34 22
> SNYYHR <- sort(NYHR)
> SNYYHR
[1] 22 25 34 35 41 41 46 46 46 47 49 54 54 59 60
> p.05 <- floor(.05*15)
> p.10 <- floor(.10*15)
> p.15 <- floor(.15*15)
> p.50 <- floor(.50*15)
> num.to.delete <-c(p.05, p.10, p.15, p.50)
> num.to.delete
[1] 0 1 2 7
> m.05 <- mean(SNYYHR[(1+p.05):(15-p.05)])
> m.10 <- mean(SNYYHR[(1+p.10):(15-p.10)])
> m.15 <- mean(SNYYHR[(1+p.15):(15-p.15)])
> m.50 <- mean(SNYYHR[(1+p.50):(15-p.50)])
> t.m <- c(m.05, m.10, m.15, m.50)
> names(t.m) <- c("5%tmean", "10%tmean", "15%tmean", "50%tmean")
> t.m
 5%tmean 10%tmean 15%tmean 50%tmean
43.93333 44.38462 44.81818 46.00000
> tm.05 <- mean(NYHR, trim=.05)
> tm.10 <- mean(NYHR, trim=.10)
> tm.15 <- mean(NYHR, trim=.15)
> tm.50 <- mean(NYHR, trim=.50)
> tms <- c(tm.05, tm.10, tm.15, tm.50)
> names(tms) <- c("5%tmean", "10%tmean", "15%tmean", "50%tmean")
> tms
 5%tmean 10%tmean 15%tmean 50%tmean
43.93333 44.38462 44.81818 46.00000
> detach(Baberuth)
```

The trimmed means are all fairly similar, confirming a rather symmetric distribution. Note that the 50% trimmed mean is the value in the middle of the sorted observations. This value is also known as the median. ■

2.5.2 The Median

While the mean is the most commonly encountered measure of center, it is not always the best measure of center. The **sample median** is the middle value of a distribution of numbers, denoted by the letter m . Since the median ignores the information in surrounding values, it is more resistant to extreme fluctuations in the data than is the mean. When working with skewed distributions, the median is the most appropriate measure of center. The sample median, m , of x_1, x_2, \dots, x_n is the $(\frac{n+1}{2})^{\text{st}}$ observation of the sorted values. When n is odd, $\frac{n+1}{2}$ is an integer, and finding the observation is straightforward. When n is even, an average of the two middle observations is taken to find the median. When the values x_1, x_2, \dots, x_n are sorted, they are called **order statistics** and denoted as $x_{(1)}, x_{(2)}, \dots, x_{(n)}$. A more concise definition of the sample median is then

$$m = \begin{cases} x_{(k+1)} & n = 2k + 1 \text{ (odd)}, \\ \frac{1}{2}(x_{(k)} + x_{(k+1)}) & n = 2k \text{ (even)}. \end{cases} \quad (2.2)$$

To find the sample median with S use the function `median(x)`, where \mathbf{x} is a numeric vector.

Example 2.11 ▷ *Means and Medians* ◁ The numerical grades achieved by three students on four exams during the course of a semester are recorded in Table 2.1. Compute means and medians for the students. Could the three students be characterized?

Table 2.1: Student test scores

	Test1	Test2	Test3	Test4
Student1	73	75	74	74
Student2	95	94	12	95
Student3	66	67	63	100

Solution: First the students exam scores are read into individual vectors denoted `Student1`, `Student2`, and `Student3`. The S function `median()` is used first to find the median test score for each student. It is possible to compute the mean test score for each student in a similar fashion to that used to find the median test score for each student. However, another solution is provided by using the S functions `rbind()`, `cbind()`, and `apply()`:

```
> Student1 <- c(73,75,74,74)
> Student2 <- c(95,94,12,95)
> Student3 <- c(66,67,63,100)
> median(Student1)
[1] 74
> median(Student2)
[1] 94.5
> median(Student3)
[1] 66.5
> SM <- rbind(Student1, Student2, Student3) # combine rows
```

```

> colnames(SM) <- c("Test1", "Test2", "Test3", "Test4")
> SM
      Test1 Test2 Test3 Test4
Student1   73   75   74   74
Student2   95   94   12   95
Student3   66   67   63  100
> means <- apply(SM, 1, mean)      # mean of rows
> medians <- apply(SM, 1, median)  # median of rows
> TOT <- cbind(SM, means, medians) # combine columns
> TOT
      Test1 Test2 Test3 Test4 means medians
Student1   73   75   74   74   74   74.0
Student2   95   94   12   95   74   94.5
Student3   66   67   63  100   74   66.5

```

As seen in the S output, the mean test score for the three students is 74. One possible characterization of the three students might be: Student 1: consistent; Student 2: overconfident; Student 3: procrastinator. Would the mean or the median be the better representative in assigning their final grades? There are good reasons one may want to consider using the median instead of the mean. ■

2.5.3 Quantiles

The p^{th} **quantile**, $0 \leq p \leq 1$, of a distribution is the value x_p such that $\mathbb{P}(X \leq x_p) \geq p$ and $\mathbb{P}(X \geq x_p) \geq 1 - p$. For discrete data, there are often many values of x_p that satisfy the definition of the p^{th} quantile. In this book, the definition used by S to compute quantiles will be used. S defines the p^{th} quantile of a distribution to be the $(p(n-1) + 1)^{\text{st}}$ order statistic. When $p(n-1) + 1$ is not an integer, linear interpolation is used between order statistics to arrive at the p^{th} quantile. Given values x_1, x_2, \dots, x_n , the p^{th} quantile for the k^{th} order statistic, $p(k)$, is

$$p(k) = \frac{(k-1)}{(n-1)}, \quad k \leq n. \quad (2.3)$$

By this definition, it is seen that the 50% quantile (50th percentile) is the median since

$$0.50 = \frac{k-1}{n-1} \Rightarrow k = \frac{n+1}{2},$$

which by definition is the location of the order statistic that is the median. Other definitions for quantiles exist and are used in other texts and other statistical software packages. However, the definition used here is consistent with S-PLUS and the default algorithm used in R for computing quantiles. To read about alternative algorithms for computing quantiles with R, type `?quantile` at the R prompt. To compute the quantiles of a data set stored in a vector `x`, use the S function `quantile(x)`. By default, the S function `quantile(x)` returns the 0%, 25%, 50%, 75%, and 100% quantiles of the data vector `x`. The p^{th} quantile is the same thing as the $(p \times 100)^{\text{th}}$ percentile. That is, percentiles and quantiles measure the same thing; however, percentiles use a scale from 0 to 100 instead of the 0 to 1 scale used by quantiles.

Just as the sample median is the value that divides the sample into equal halves, the **sample quartiles** can be thought of as the values that divide the sample into quarters. The first, second, and third sample quartiles are denoted as Q_1 , Q_2 , and Q_3 , respectively, and are (by default) computed with the S function `quantile(x)`. To compute other quantiles,

use the argument `probs=` to specify either a single value or to pass a vector of values to the `quantile()` function.

Example 2.12 Compute Q_1 , Q_2 , and Q_3 for the values $x_{(1)} = 1$, $x_{(2)} = 4$, $x_{(3)} = 7$, $x_{(4)} = 9$, $x_{(5)} = 10$, $x_{(6)} = 14$, $x_{(7)} = 15$, $x_{(8)} = 16$, $x_{(9)} = 20$, and $x_{(10)} = 21$.

Solution: First, the order statistics for the 0.25, 0.50, and 0.75 quantiles are computed using (2.3):

$$\begin{array}{lll} .25 = \frac{k-1}{10-1} & .50 = \frac{k-1}{10-1} & .75 = \frac{k-1}{10-1} \\ k = 3.25 & k = 5.50 & k = 7.75 \end{array}$$

Linear interpolation is then used on the order statistics to find the requested quantiles/quartiles. Specifically, since Q_1 , Q_2 , and Q_3 occur at the 3.25, 5.50, and 7.75 order statistics, 0.25 of the distance between the third and fourth order statistics is added to the third order statistic to arrive at Q_1 . Likewise, 0.50 of the distance between the fifth and sixth order statistics is added to the fifth order statistic to compute Q_2 . Finally, 0.75 of the distance between the seventh and eighth order statistics is added to the seventh order statistic to compute Q_3 :

$$\begin{array}{ll} Q_1 = x_{(3)} + .25(x_{(4)} - x_{(3)}) & Q_2 = x_{(5)} + .50(x_{(6)} - x_{(5)}) \\ = 7 + .25(9 - 7) & = 10 + .5(14 - 10) \\ = 7.50 & = 12.00 \\ \\ Q_3 = x_{(7)} + .75(x_{(8)} - x_{(7)}) & \\ = 15 + .75(16 - 15) & \\ = 15.75 & \end{array}$$

Code to compute the requested quartiles according to the quantile definition follows. Subsequently, the S function `quantile()` is used to compute the same quantiles/quartiles.

```
> x <- c(1,4,7,9,10,14,15,16,20,21)
> p <- c(.25,.5,.75)           # desired quantiles
> n <- length(x)              # number of values, n
> order.stat <- p*(n-1)+1     # computing order statistics
> order.stat                   # order statistics
[1] 3.25 5.50 7.75
> Q1 <- x[3]+.25*(x[4]-x[3])   # linear interpolation
> Q2 <- x[5]+.50*(x[6]-x[5])   # linear interpolation
> Q3 <- x[7]+.75*(x[8]-x[7])   # linear interpolation
> QU <- c(Q1, Q2, Q3)
> names(QU) <- c("Q1","Q2","Q3")
> QU                            # quartiles
  Q1   Q2   Q3
 7.50 12.00 15.75
> quantile(x, probs=c(.25,.5,.75)) # the easy way!
 25%  50%  75%
 7.50 12.00 15.75
```



2.5.4 Hinges and Five-Number Summary

An alternative method to calculating quartiles is to compute **hinges**. The idea behind both quartiles and hinges is to split the data into fourths. When a computer is not available, hinges are somewhat easier to compute by hand than are quartiles. The lower and upper hinges are the $x_{(j)}$ and $x_{(n-j+1)}$ order statistics, where

$$j = \frac{\lfloor \frac{n+1}{2} \rfloor + 1}{2}. \quad (2.4)$$

In short, the lower hinge is the median of the lower half of the data and the upper hinge is the median of the upper half of the data. Lower and upper hinges can be different from quartiles. For example, consider Example 2.12 on the previous page where the locations of the first, and third quartiles were found to be at the 3.25th and 7.75th order statistics. However, since

$$\frac{\lfloor \frac{n+1}{2} \rfloor + 1}{2} = \frac{\lfloor \frac{10+1}{2} \rfloor + 1}{2} = 3,$$

the locations for the lower and upper hinges are at the 3rd, $x_{(3)}$, and 8th, $x_{(n-3+1)} = x_{(10-3+1)} = x_{(8)}$, order statistics.

Hinges are typically returned as part of the **five-number summary**. A five-number summary for a data set consists of the smallest value, the lower hinge, the median, the upper hinge, and the largest value, all of which are computed with R's function `fivenum()`.

Example 2.13 Compute the 0.25, 0.50, and 0.75 quantiles as well as a five-number summary for the number of runs batted in (RBIs) by Babe Ruth while he played for the New York Yankees. The variable `RBI` in the data frame `Baberuth` contains the RBIs per season for Babe Ruth over his professional baseball career.

Solution: The quartiles and hinges are first computed by their definitions. Subsequently, the S function `quantile()` and the R function `fivenum()` are used to obtain the same results:

```
> attach(Baberuth)      # Assumes package PASWR is loaded
> NYRBI <- RBI[7:21]    # Extract RBIs only while a NYY
> SNYYRBI <- sort(NYYRBI)
> p <- c(.25, .50, .75)
> n <- length(NYYRBI)
> n
[1] 15
> order.stat <- p*(n-1)+1
> order.stat
[1] 4.5 8.0 11.5
> Q1 <- SNYYRBI[4] + .5*(SNYYRBI[5] - SNYYRBI[4])
> Q2 <- SNYYRBI[8]
> Q3 <- SNYYRBI[11] + .5*(SNYYRBI[12] - SNYYRBI[11])
> QU <- c(Q1, Q2, Q3)
> names(QU) <- c("Q1", "Q2", "Q3")
> QU
      Q1      Q2      Q3
112.0 137.0 153.5
```

```

> quantile(NYYRBI, probs=c(.25,.50,.75))
  25%  50%  75%
112.0 137.0 153.5
> j <- (floor((n+1)/2)+1)/2      # Number to count in
> j
[1] 4.5
> lower.hinge <- SNYYRBI[4]+.5*(SNYYRBI[5]-SNYYRBI[4])
> upper.hinge <- SNYYRBI[11]+.5*(SNYYRBI[12]-SNYYRBI[11])
> small <- min(NYYRBI)
> large <- max(NYYRBI)
> five.numbers <- c(small, lower.hinge, Q2, upper.hinge, large)
> five.numbers
[1]  66.0 112.0 137.0 153.5 171.0
> fivenum(NYYRBI)                # Only works in R
[1]  66.0 112.0 137.0 153.5 171.0
> detach(Baberuth)

```

In this particular example, the first and third quartile are equal to the lower and upper hinge, respectively. ■

2.5.5 Boxplots

A popular method of representing the information in a five-number summary is the **boxplot**. To show spread, a box is drawn from the lower hinge (H_L) to the upper hinge (H_U) with a vertical line drawn through the box to indicate the median or second quartile (Q_2). A “whisker” is drawn from H_U to the largest data value that does not exceed the upper fence. This value is called the **adjacent value**. The upper fence is defined as $Fence_U = H_U + 1.5 \times H_{spread}$, where $H_{spread} = H_U - H_L$. A whisker is also drawn from H_L to the smallest value that is larger than the lower fence, where the lower fence is defined as $Fence_L = H_L - 1.5 \times H_{spread}$. Any value smaller than the lower fence or larger than the upper fence is considered an **outlier** and is generally depicted with a hollow circle. Figure 2.10 on the following page illustrates a boxplot for the variable **fat** from the data frame **Bodyfat**.

To create a boxplot with S, use the command `boxplot()`. By default, boxplots in R have a vertical orientation. To create a horizontal boxplot with R, use the optional argument `horizontal=TRUE`. Currently, S-PLUS does not have an option to produce horizontal boxplots with the `boxplot()` function. However, S-PLUS does have the function `bwplot()`, which produces horizontal boxplots. Common arguments for `boxplot()` include `col=` to set the box color and `notch=TRUE` to add a notch to the box to highlight the median.

Example 2.14 Use the data frame **Cars93** in the MASS package to create a boxplot of the variable **Min.Price**. Use the `text()` function to label the five-number summary values in the boxplot.

Solution: Two solutions are presented: one for R and one for S-PLUS. The solution for S-PLUS is slightly more involved because S-PLUS does not have a built-in function to compute the five-number summary. The final boxplot from R is shown in Figure 2.11 on page 47. Additionally, the labels in R contain mathematical notation. To learn more about R’s ability to plot mathematical expressions, type `?mathplot` at the R prompt.

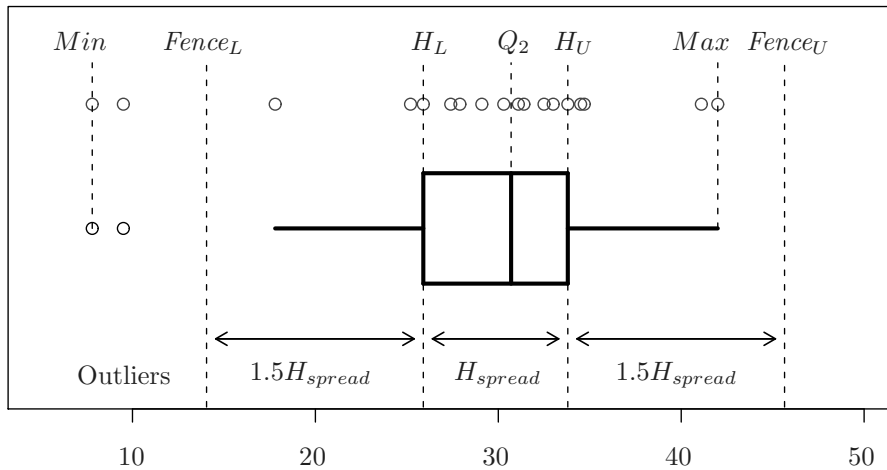


FIGURE 2.10: Graph depicting the five-number summary in relationship to original data and the boxplot

Solution for R:

```
> library(MASS)
> attach(Cars93)
> boxplot(Min.Price, ylab="Minimum Price (in $1000) for basic
+ version", col="gray")
> f <- fivenum(Min.Price)
> text(rep(1.25,5), f, labels=c("Min", expression(H[L]),
+ expression(Q[2]), expression(H[U]), "Max"), pos=4)
> detach(Cars93) # Clean up
```

Solution for S-PLUS:

```
> library(MASS)
> attach(Cars93)
> n <- length(Min.Price)
> smp <- sort(Min.Price)
> count <- (floor((n+1)/2)+1)/2 # Using Equation 2.4
> count
[1] 24
> lower.hinge <- smp[count]
> upper.hinge <- smp[(n-count+1)]
> five.num <- c(min(smp), lower.hinge, median(smp), upper.hinge,
+ max(smp))
> boxplot(Min.Price, ylab="Minimum Price (in $1000) for basic
+ version")
> text(rep(85,5), five.num, labels=c("Minimum", "Lower Hinge",
+ "Median", "Upper Hinge", "Maximum"))
> detach(Cars93) # Clean up
```

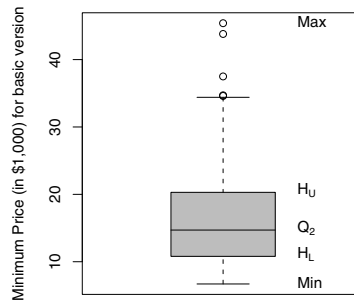


FIGURE 2.11: Boxplot of car prices with five-number summaries labeled

Boxplots are useful for detecting skewness, finding outliers, and comparing two or more variables that are all measured on the same scale. However, a boxplot will not detect multi-modality. ■

2.6 Summary Measures of Spread

Summary measures of center such as the mean and median are important because they describe “typical” values in a data set. However, it is possible to have two data sets with the same means (Example 2.11 on page 41) and/or medians while still having different spreads. For this reason, it is important to measure not only typical values but also the spread of the values in a distribution in order to describe the distribution fully. There are many ways to measure spread, some of which include range, interquartile range, and variance.

2.6.1 Range

The easiest measure of spread to compute is the range. At times, the range refers to the difference between the smallest value in a data set and the largest value in the data set. Other times, the range refers to the smallest and largest values of a data set as a pair. The S function `range(x)` returns the smallest and largest values in `x`. If the distance between the largest and smallest value is desired, one can use `diff(range(x))`:

```
> range(1:10)
[1] 1 10
> diff(range(1:10))
[1] 9
```

2.6.2 Interquartile Range

Instead of looking at the entire range of the data, looking at the middle 50% will often prove to be a useful measure of spread, especially when the data are skewed. The interquartile range (*IQR*) is defined as $IQR = Q_3 - Q_1$ and can be found with the function `IQR()`:

```
> quantile(1:10)
  0%   25%   50%   75%  100%
 1.00  3.25  5.50  7.75 10.00
> IQR(1:10)
[1] 4.5
```

2.6.3 Variance

The **sample variance**, s^2 , can be thought of as the average squared distance of the sample values from the sample mean. It is not quite the average because the quantity is divided by $n - 1$ instead of n in the formula

$$s^2 = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n - 1}. \quad (2.5)$$

When the positive square root of the sample variance is taken, the **sample standard deviation**, s , results. It is often preferable to report the sample standard deviation instead of the variance since the units of measurement for the sample standard deviation are the same as those of the individual data points in the sample. To compute the variance with S, use the function `var(x)`. One could compute the standard deviation by taking the square root of the variance `sqrt(var(x))` or use the built-in function to do so. However, be aware that the function to compute the standard deviation in R is `sd(x)`, while the function to compute the standard deviation in S-PLUS is `stdev(x)`. The standard deviation is an appropriate measure of spread for normal distributions:

```
> x <- 1:5
> x
[1] 1 2 3 4 5
> n <- length(x)
> mean.x <- mean(x)
> mean.x
[1] 3
> x-mean.x
[1] -2 -1 0 1 2
> (x-mean.x)^2
[1] 4 1 0 1 4
> NUM <- sum((x-mean.x)^2) # numerator of s^2 hard way
> NUM
[1] 10
> DEN <- n-1 # denominator of s^2
> DEN
[1] 4
> VAR <- NUM/DEN # variance hard way
> VAR
[1] 2.5
> var(x) # variance easy way
[1] 2.5
> SD <- sqrt(VAR) # standard deviation hard way
> SD
[1] 1.581139
```

```

> sd(x)                # standard deviation with R
[1] 1.581139
> stdev(x)             # standard deviation with S-PLUS
[1] 1.581139

```

An interesting function that will return different results depending on the class of the object to which it is applied is the S function `summary()`. When the object is a numeric vector, as is the case with `x`, six summary statistics are returned: the minimum, the first quartile, the median, the mean, the third quartile, and the maximum:

```

> summary(x)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    1         2         3         3         4         5

```

2.7 Bivariate Data

Methods to summarize and display relationships between two variables (bivariate data) will be the focus of the next few pages. In Section 2.3, two of the methods used to gain a deeper understanding of categorical variables were tables and barplots. When two variables are categorical, tables, called contingency tables, and barcharts will still prove useful. When presented with quantitative bivariate data, relevant questions will deal with the relationships between the two variables. For example, is there a relationship between a person's height and his weight? Is there a relationship between the amount of time a student spends studying and her grades? Graphical techniques such as scatterplots can be used to explore bivariate relationships. When relationships exist between variables, different correlation coefficients are used to characterize the strengths of the relationships. Finally, a brief introduction to the simple linear regression model is given before moving into multivariate data.

2.7.1 Two-Way Contingency Tables

The S command `table(x)` was used for creating frequency tables with univariate, categorical variables. For bivariate, categorical data, the command `table(x, y)` is used to create two-way contingency tables where `x` and `y` represent the two categorical variables.

Example 2.15 Consider the data frame `EPIDURAL`, which contains information from a study to determine whether the traditional sitting position or the hamstring stretch position is superior for administering epidural anesthesia to pregnant women in labor as measured by the number of obstructive (needle to bone) contacts (OC). The variable `Doctor` specifies which of the four physicians in the study administered the procedure. `Ease` is the physician's assessment prior to administering the epidural of how well bony landmarks for epidural placement can be felt. Produce a two-way contingency table for the variables `Doctor` and `Ease`.

Solution: The goal is to produce a two-way table such as the one in Table 2.2 on the next page with S. The levels of categorical variables by default are alphabetical. Consequently, the levels of `Ease` are Difficult, Easy, and Impossible. Pay particular attention to how the levels of a variable can be rearranged in the code that follows.

Table 2.2: Two-way table of Doctor by Ease

	Easy	Difficult	Impossible
Dr. A	19	3	1
Dr. B	7	10	4
Dr. C	18	3	0
Dr. D	13	4	3

```

> head(EPIDURAL, n=5)      # Shows first five rows of EPIDURAL
  Doctor kg cm Ease Treatment OC Complications
1 Dr. B 116 172 Difficult Traditional Sitting 0 None
2 Dr. C 86 176 Easy Hamstring Stretch 0 None
3 Dr. B 72 157 Difficult Traditional Sitting 0 None
4 Dr. B 63 169 Easy Hamstring Stretch 2 None
5 Dr. B 114 163 Impossible Traditional Sitting 0 None
> attach(EPIDURAL)
> table(Doctor, Ease)      # Levels of Ease not in increasing order
      Ease
Doctor Difficult Easy Impossible
  Dr. A          3  19           1
  Dr. B         10   7           4
  Dr. C          3  18           0
  Dr. D          4  13           3
> Teasy <- factor(Ease, levels=c("Easy","Difficult","Impossible"))
> table(Doctor, Teasy)     # Levels of Ease in increasing order
      Teasy
Doctor Easy Difficult Impossible
  Dr. A  19          3           1
  Dr. B   7         10           4
  Dr. C  18          3           0
  Dr. D  13          4           3

```

Although the command `table(Doctor, Ease)` produced a two way contingency table, reordering the levels of `Ease` produces a more readable two-way contingency table. ■

Extensions to multi-way contingency tables can be accomplished by specifying additional factors to the `table()` function or by using the R flattened table function `ftable()`. A flattened three-way contingency table of the factors `Doctor`, `Treatment`, and `Teasy` follows. More options for both `table` and `ftable` can be found in their respective help files.

```

> ftable(Doctor, Treatment, Teasy)
      Teasy Easy Difficult Impossible
Doctor Treatment
Dr. A Hamstring Stretch          7           1           0
      Traditional Sitting       12           2           1
Dr. B Hamstring Stretch          3           3           0
      Traditional Sitting         4           7           4
Dr. C Hamstring Stretch          8           3           0
      Traditional Sitting       10           0           0
Dr. D Hamstring Stretch          7           1           2
      Traditional Sitting         6           3           1
> detach(EPIDURAL)

```

2.7.2 Graphical Representations of Two-Way Contingency Tables

Barplots can be used to depict graphically the information from two-way contingency tables. This is accomplished by picking one of the variables to form the categories of the barplot. Next, the second variable's levels are graphed either in a single bar (stacked) or as several bars (side-by-side).

Example 2.16 Produce stacked and side-by-side barplots of the information contained in Table 2.2 on the facing page.

Solution: Barplots where the variable of interest is `Ease` then `Doctor` are created first. Subsequently, side-by-side barplots where the variables of interest are `Ease` then `Doctor` are created. The graphs in Figure 2.12 on the next page were created using R. Output from S-PLUS will look slightly different. The user should consult the on-line documentation using `?barplots` for the differences between R and S-PLUS.

```
> attach(EPIDURAL)
> Teasy <- factor(Ease, levels=c("Easy","Difficult","Impossible"))
> X <- table(Doctor, Teasy)
> X
```

		Teasy		
Doctor		Easy	Difficult	Impossible
Dr. A	19	3	1	
Dr. B	7	10	4	
Dr. C	18	3	0	
Dr. D	13	4	3	

```
> t(X)      # Transpose X
      Doctor
Teasy  Dr. A Dr. B Dr. C Dr. D
Easy   19   7   18   13
Difficult 3  10   3   4
Impossible 1   4   0   3
```

```
> par(mfrow=c(2,2))
> barplot(X, main="Barplot where Doctor is Stacked \n within Levels
+ of Palpitation")
> barplot(t(X), main="Barplot where Levels of Palpitation \n is
+ Stacked within Doctor")
> barplot(X, beside=TRUE, main="Barplot where Doctor is Grouped \n
+ within Levels of Palpitation")
> barplot(t(X), beside=TRUE, main="Barplot where Levels of Palpitation
+ \n is Grouped within Doctor")
> par(mfrow=c(1,1))
> detach(EPIDURAL)
```

From the example, it is seen that the categories for the barplot are the numeric columns in a two-way contingency table. If the user wants the categories to be reversed, transpose the table using the command `t(table(x, y))`, where `table(x, y)` is the two-way contingency table. ■

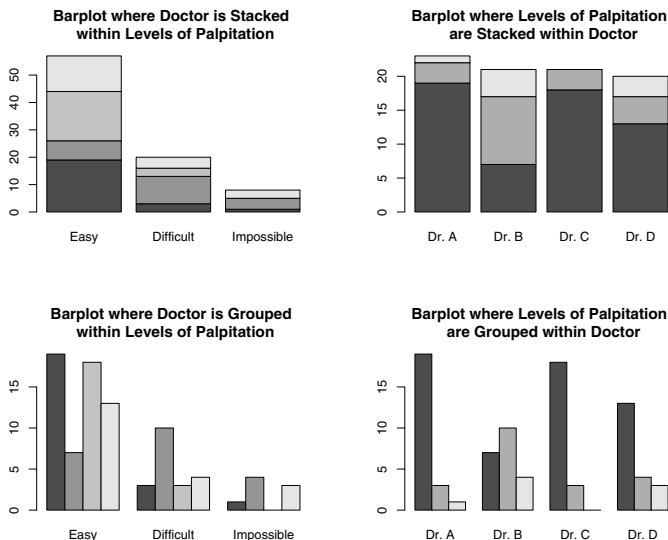


FIGURE 2.12: Stacked and side-by-side barplots for levels of palpitation (Teasy) and physician (Doctor)

Relationships are often better represented with proportions than with counts. R has the function `prop.table(x)`, which can be used to compute proportions based on the number of entries in either the entire table, `x`, which is the default, or by entering `prop.table(x, margin=1)` for row totals or `prop.table(x, margin=2)` for column totals.

Example 2.17 Using the data frame `EPIDURAL`, create a side-by-side barplot of `Treatment` versus `OC`.

Solution: Since there have been 25 patients treated with the hamstring stretch position and 49 patients treated with the traditional sitting position, it would not be rational to compare the frequencies. Instead, one should compare the percentages within the categories of `OC` by `Treatment`:

```
> attach(EPIDURAL)
> table(Treatment, OC)
          OC
Treatment 0  1  2  3  4  5  6 10
Hamstring Stretch 17  6  6  2  1  1  0  2
Traditional Sitting 23 16  3  1  2  2  2  0
> addmargins(table(Treatment, OC)) # addmargins is an R command
          OC
Treatment 0  1  2  3  4  5  6 10 Sum
Hamstring Stretch 17  6  6  2  1  1  0  2 35
Traditional Sitting 23 16  3  1  2  2  2  0 49
Sum              40 22  9  3  3  3  2  2 84
> X <-prop.table(table(Treatment, OC),1) # Percents by rows
```

```

> X
              OC
Treatment      0      1      2      3
Hamstring Stretch 0.48571429 0.17142857 0.17142857 0.05714286
Traditional Sitting 0.46938776 0.32653061 0.06122449 0.02040816
              OC
Treatment      4      5      6      10
Hamstring Stretch 0.02857143 0.02857143 0.00000000 0.05714286
Traditional Sitting 0.04081633 0.04081633 0.04081633 0.00000000

> par(mfrow=c(2,1))
> barplot(X, beside=TRUE, legend=TRUE)
> barplot(t(X), beside=TRUE, legend=TRUE)
> par(mfrow=c(1,1))
> detach(EPIDURAL)

```

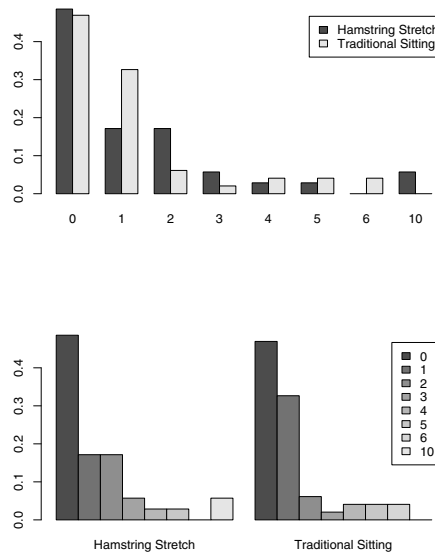


FIGURE 2.13: Side-by-side barplots showing percentages in obstructive contacts categories by treatments

Note that the categories for the barplot in the upper graph of Figure 2.13 are the OC categories in the two-way contingency table. Within each OC category, comparisons are shown side-by-side based on the treatment. If the user wants the categories to be reversed, transpose the table using the command `t(table(x, y))`, where `table(x, y)` is the two-way contingency table. ■

2.7.3 Comparing Samples

The need to compare two samples is quite common. Simple experiments will often compare a control group to a treatment group in an effort to see if the treatment provides

some added benefit. For example, the data in the `EPIDURAL` data frame are from an ongoing experiment to see which of two positions results in fewer obstructive bone contacts (the times the needle hits a bone). When comparing two samples, typically some type of inference to the samples' populations is desired. That is, are the centers the same? Are the spreads similar? Are the shapes of the two distributions similar? Graphs such as histograms, density plots, boxplots, and quantile-quantile plots can help answer these questions. Histograms and density plots were introduced in Section 2.4, and boxplots were introduced in Section 2.5. A **quantile-quantile (Q-Q) plot** plots the quantiles of one distribution against the quantiles of another distribution as (x, y) points. When the two distributions have similar shapes, the points will fall along a straight line. The `S` function to make a quantile-quantile plot is `qqplot(x, y)`. Histograms can be used to compare two distributions. However, it is rather challenging to put both histograms on the same graph. Example 2.18 shows the user how histograms can be used to compare distributions. However, a better approach is to use Trellis/lattice graphics, which are explained in Section 2.8.

Example 2.18 Use histograms to compare the body weight index (BWI) for the two treatments (traditional sitting and hamstring stretch stored in `Treatment`) using the data frame `EPIDURAL`.

Solution: First, BWI is typically defined as kg/m^2 . Since the data frame `EPIDURAL` does not contain a BWI variable, one is created. Subsequently, the default options for the BWI histograms of the control and treatment groups are shown in the first column of Figure 2.14 on the next page, while the BWI histograms of the control and treatment groups are shown in the second column of Figure 2.14 after the axes limits for both the x - and y -axes have been set to the same values for both histograms:

```
> attach(EPIDURAL)
> BWI <- kg/(cm/100)^2
> Control <- BWI[Treatment=="Traditional Sitting"]
> Treated <- BWI[Treatment=="Hamstring Stretch"]
> par(mfrow=c(2,2))      # 2*2 plotting region
> hist(Control)
> hist(Control, xlim=c(20,60), ylim=c(0,17))
> hist(Treated)
> hist(Treated, xlim=c(20,60), ylim=c(0,17))
> par(mfrow=c(1,1))      # 1*1 plotting region
> detach(EPIDURAL)
```

Note that it is misleading to compare histograms where the bin widths and/or units on the axes of the two histograms are different. Note that both axes are different in the first column of Figure 2.14 on the facing page. The bins of the two histograms are set with the argument `breaks=`, and the x - and y -axes are set with the arguments `xlim=` and `ylim=`, respectively. The general shape of the BWI for the patients administered epidurals in the hamstring stretch position is unimodal skewed to the right. While the distribution of BWI for patients administered epidurals in the traditional sitting position is also unimodal skewed to the right, it is not quite as skewed as the distribution where patients are administered epidurals from the hamstring stretch position. ■

Example 2.19 Use side-by-side boxplots and superimposed density plots to compare the BWI for the two treatments (traditional sitting and hamstring stretch stored in `Treatment`) using the data frame `EPIDURAL`.

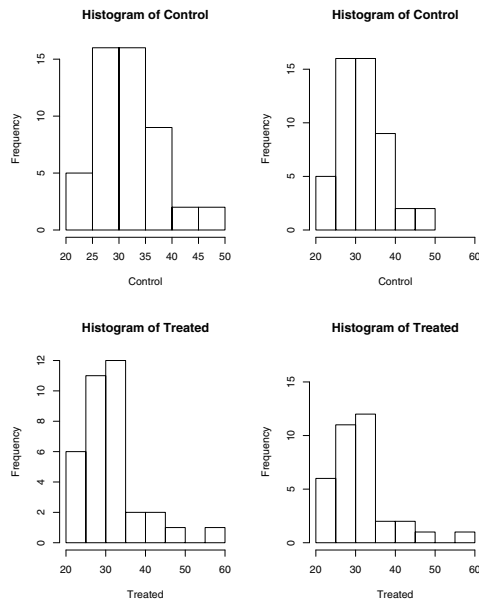


FIGURE 2.14: Side-by-side barplots showing percentages in obstructive contacts' categories by treatments

Solution: The argument `horizontal=TRUE` used in the `boxplot()` function will only work in R. One way to create horizontal boxplots with S-PLUS is to use the Trellis function `bwplot()`. Specifically, one might enter `bwplot(Treatment~BWI)` to produce side-by-side boxplots with S-PLUS. Trellis/lattice graphs will be discussed in more detail in Section 2.8. Using boxplots, as seen in Figure 2.15 on the next page, one sees that the median for both treatments is around 30 kg/m^2 and both distributions appear to be skewed to the right.

```
> attach(EPIDURAL)
> par(pty="s")          # Make plotting region square
> BWI <- kg/(cm/100)^2 # Define body weight index
> Control <- BWI[Treatment=="Traditional Sitting"]
> Treated <- BWI[Treatment=="Hamstring Stretch"]
> boxplot(Control, Treated, horizontal=TRUE, col=c(13,4),
+ names=c("Traditional Sitting","Hamstring Stretch"), las=1)
> plot(density(Control), xlim=c(20,60), col=13, lwd=2, main="", xlab="")
> lines(density(Treated), lty=2, col=4, lwd=2)
> detach(EPIDURAL)
```

The density plot in Figure 2.16 on the following page further indicates that the distributions for the BWI for both the traditional sitting and the hamstring stretch position are skewed to the right. ■

Example 2.20 Use a quantile-quantile plot to compare the BWI for the two treatments (traditional sitting and hamstring stretch stored in `Treatment`) using the data frame `EPIDURAL`.

Solution: Commands to recreate the quantile-quantile plot shown in Figure 2.17 on page 57 follow. Note that both the x - and y -axes have the same limits.

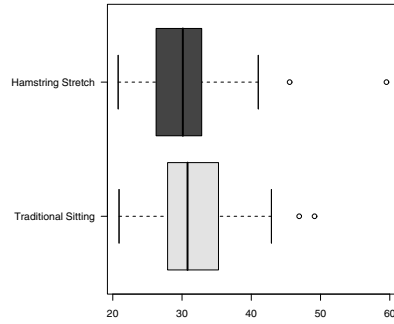


FIGURE 2.15: Side-by-side boxplots of BWI in the traditional sitting and hamstring stretch positions

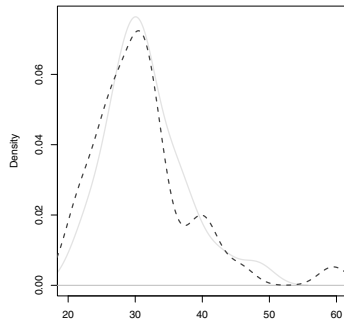


FIGURE 2.16: Density plots of BWI in the traditional sitting (solid line) and hamstring stretch positions (dashed line)

```
> attach(EPIDURAL)
> par(pty="s")      # Make plotting region square
> qqplot(Control, Treated, xlim=c(20,60), ylim=c(20,60))
> abline(a=0, b=1) # Line y=0+1*x
> par(pty="m")     # Maximize plotting region
> detach(EPIDURAL)
```

The quantile-quantile plot in Figure 2.17 suggests the distributions are fairly similar since the points roughly follow the $y = x$ line. ■

2.7.4 Relationships between Two Numeric Variables

Relationships between two numeric variables can be viewed with **scatterplots**. A scatterplot plots the values of one variable against the values of a second variable as points (x_i, y_i) in the Cartesian plane. Typical questions researchers seek to answer with numeric variables include “Is there a relationship between the two variables?”, “How strong is the relationship between the two variables?”, and “Is the relationship linear?” Questions such as

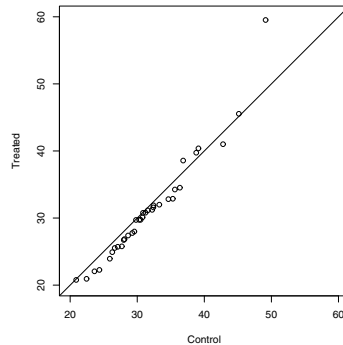


FIGURE 2.17: Quantile-quantile plot of BWI in the traditional sitting and hamstring stretch positions

“Is there a relationship between a person’s height and his weight?” or “Is there a relationship between a student’s grades and the time spent studying?” are typical. Given two numeric variables, say x and y , entering the S function `plot(x, y)` produces a scatterplot.

Example 2.21 Use the data frame `Animals` from the `MASS` package to investigate whether the brain weights of animals are related to their body weights. In other words, is a bigger brain required to govern a bigger body?

Solution: Because of the large range in body and brain weights, (0.023 kg to 87,000 kg) and (0.4 g to 5712 g), respectively, a scatterplot of the values in `body` and `brain` is too distorted to reveal any clear pattern. Consequently, the data is transformed by taking natural logarithms of both variables and plotting the resulting values as shown in Figure 2.18 on the following page.

```
> library(MASS)
> attach(Animals)
> range(body)
[1] 2.3e-02 8.7e+04
> range(brain)
[1] 0.4 5712.0
> range(log(body))
[1] -3.772261 11.373663
> range(log(brain))
[1] -0.9162907 8.6503245
> par(pty="s")
> plot(log(body), log(brain))
> identify(log(body), log(brain), labels=row.names(Animals))
> detach(Animals)
```

The function `identify()` was used to label several of the points in Figure 2.18 on the next page. The function `identify()` labels the closest point in the scatterplot with each mouse click (left click with windows) until instructed to stop. How the function is instructed to stop varies by operating system. Right clicking with windows, middle clicking with Linux, and using the escape key in Mac OS X will generally stop the identification process. Based on Figure 2.18, there appears to be linear relationship between the logarithm of the body weights and the logarithm of the brain weights. The dinosaurs can be classified as bivariate outliers as they do not fit the overall pattern seen in the rest of the data. ■

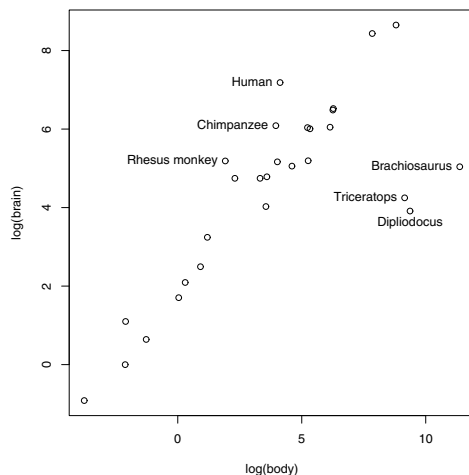


FIGURE 2.18: Scatterplot of $\log(\text{brain})$ versus $\log(\text{body})$ for Example 2.21

2.7.5 Correlation

The **correlation coefficient**, denoted by r , measures the strength and direction of the linear relationship between two numeric variables X and Y and is defined by

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_X} \right) \left(\frac{y_i - \bar{y}}{s_Y} \right) \quad (2.6)$$

The value for r will always be between -1 and $+1$. When r is close to $+1$, it indicates a strong positive linear relationship. That is, when x increases so does y , and vice versa. When the value of r is close to -1 , it indicates a strong negative linear relationship. Values of r close to zero indicate weak linear relationships. To compute the correlation between two numeric vectors with **S**, one may use the function `cor(x, y)`.

Example 2.22 Find the correlation coefficient, r , between the logarithms of the body and brain weights in the data frame `Animals` from the `MASS` package using (2.6). Verify the calculated answer using the **S** function `cor()`.

Solution: First, the variables `logbody`, `logbrain`, `Zbody`, and `Zbrain` are created. The new variables are subsequently column binded to the `Animals` data frame and stored in a new data frame named `Anim`. Note that the output uses the **R** function `sd()` to compute the standard deviation. To compute the standard deviation with **S-PLUS**, use `stdev()`.

```
> attach(Animals)
> options(digits=3)           # Three digits for output
> logbody <- log(body)
> logbrain <- log(brain)
> Zbody <- (logbody - mean(logbody))/sd(logbody)
> Zbrain <- (logbrain - mean(logbrain))/sd(logbrain)
> Anim <- cbind(Animals, logbody, logbrain, Zbody, Zbrain)
> n <- length(logbody)
```

```

> r <- (1/(n-1))*sum(Zbody*Zbrain) # Definition of r
> r
[1] 0.78
> cor(logbody, logbrain)
[1] 0.78
> Anim[1:6,]
# Show first 6 rows of Anim
      body  brain logbody logbrain  Zbody  Zbrain
Mountain beaver  1.35e+00   8.1  0.3001   2.092 -0.9206 -0.9726
Cow              4.65e+02  423.0  6.1420   6.047  0.6287  0.6760
Grey wolf        3.63e+01  119.5  3.5926   4.783 -0.0474  0.1492
Goat             2.77e+01  115.0  3.3200   4.745 -0.1197  0.1332
Guinea pig      1.04e+00   5.5  0.0392   1.705 -0.9898 -1.1340
Dipliodocus     1.17e+04  50.0  9.3673   3.912  1.4841 -0.2140

```

The correlation between `logbrain` and `logbody` is 0.78, which indicates a positive linear relationship between the two variables. An alternative to computing the z-scores directly is to use the function `scale()`:

```

> ZBO <- scale(logbody) # Z score of logbody
> ZBR <- scale(logbrain) # Z score of logbrain
> SAME <- cbind(Zbody, ZBO, Zbrain, ZBR)
> SAME[1:5,]
# Show first five rows of data frame
      Zbody      Zbrain
[1,] -0.9206 -0.9206 -0.973 -0.973
[2,]  0.6287  0.6287  0.676  0.676
[3,] -0.0474 -0.0474  0.149  0.149
[4,] -0.1197 -0.1197  0.133  0.133
[5,] -0.9898 -0.9898 -1.134 -1.134
> detach(Animals)

```

2.7.6 Sorting a Data Frame by One or More of Its Columns

The `sort()` function can be used to sort a single variable in either increasing or decreasing order. However, if the user wants to sort a variable in a data frame and have the other variables reflect the new ordering, `sort()` will not work. The function needed to rearrange the values in a data frame to reflect the order of a particular variable or variables in the event of ties is `order()`. Given three variables `x`, `y`, and `z` in a data frame `DF`, the command `order(x)` returns the indices of the sorted values of `x`. Consequently, the data frame `DF` can be sorted by `x` with the command `DF[order(x),]`. In the event of ties, further arguments to `order` can be used to specify how the ties should be broken. Consider how ties are broken with the following numbers. To conserve space, the transpose function `t()` is used on the data frame `DF`.

```

> x <- c(1,1,1,3,3,3,2,2,2)
> y <- c(3,2,3,6,2,6,10,4,4)
> z <- c(7,4,2,9,6,4,5,3,1)
> DF <- data.frame(x, y, z)
> rm(x, y, z)
> attach(DF)

```

```

> t(DF)
  1 2 3 4 5 6  7 8 9
x 1 1 1 3 3 3  2 2 2
y 3 2 3 6 2 6 10 4 4
z 7 4 2 9 6 4  5 3 1
> t(DF[order(x, y, z),])
  2 3 1 9 8  7 5 6 4
x 1 1 1 2 2  2 3 3 3
y 2 3 3 4 4 10 2 6 6
z 4 2 7 1 3  5 6 4 9
> detach(DF)

```

Example 2.23 Find the correlation coefficient, r , between the logarithms of the body and brain weights in the data frame `Animals` from the `MASS` package with and without dinosaurs.

Solution: To save space, only four rows of the data frames `SA` and `NoDINO` are shown in the output. Note that there are a total of 28 animals in the data frame `Animals`.

```

> attach(Animals)
> cor(log(body), log(brain))
[1] 0.7794935
> SA <- Animals[order(body),] # Sorted by body weight
> detach(Animals)
> tail(SA, n=4) # Equivalently SA[25:28,], shows four heaviest animals
      body brain
African elephant 6654 5712.0
Triceratops      9400  70.0
Dipliodocus      11700  50.0
Brachiosaurus    87000 154.5
> NoDINO <- SA[-(28:26),] # Remove rows 26-28 of SA
> attach(NoDINO)          # NoDINO contains 25 rows
> NoDINO[22:25,]         # Show four heaviest animals
      body brain
Horse      521  655
Giraffe    529  680
Asian elephant 2547 4603
African elephant 6654 5712
> cor(log(body), log(brain)) # Correlation without dinosaurs
[1] 0.9600516
> detach(NoDINO)

```

The correlation between $\log(\text{brain})$ and $\log(\text{body})$ when dinosaurs are included is 0.78 and the correlation between $\log(\text{brain})$ and $\log(\text{body})$ is 0.96 when the dinosaurs are removed from the computation. ■

2.7.7 Fitting Lines to Bivariate Data

When a linear pattern is evident from a scatterplot, the relationship between the two variables is often modeled with a straight line. When modeling a bivariate relationship, Y is called the **response** or **dependent** variable, and x is called the **predictor** or **independent** variable. There are relationships that are of interest that are not linear. However, before

addressing more complicated models, this material attempts to provide a foundation for the simpler models (simple linear regression) from which more complicated models can later be built. Chapter 12 is devoted to standard regression techniques for both the simple and multiple linear regression model. The simple linear regression model is written

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (2.7)$$

Model (2.7) is said to be simple, linear in the parameters (β_0 and β_1), and linear in the predictor variable (x_i). It is simple because there is only one predictor; linear in the parameters because no parameter appears as an exponent nor is multiplied or divided by another parameter; and linear in the predictor variable since the predictor variable is raised only to the first power. When the predictor variable is raised to a power, this power is called the **order** of the model. For now, only the simple linear model will be discussed. The goal is to estimate the coefficients β_0 and β_1 in (2.7). The most well-known method of estimating the coefficients β_0 and β_1 is to use ordinary least squares (OLS). OLS provides estimates of β_0 and β_1 by minimizing the sum of the squared deviations of the Y_i s for all possible lines. Specifically, the sum of the squared residuals ($\hat{\varepsilon}_i = Y_i - \hat{Y}_i$) is minimized when the OLS estimators of β_0 and β_1 are

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \quad (2.8)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad (2.9)$$

respectively. Note that the estimated regression function is written as

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

A graphical representation of the residuals and a line fit to some data using OLS can be seen in Figure 2.19 on the following page.

The OLS estimators of β_0 and β_1 are affected by outliers just as the mean and standard deviation are subject to outliers. Recall that the median and IQR were suggested as measures of center and spread, respectively, when working with skewed distributions. This recommendation was made because the median and IQR provide more robust measures of center and spread in the presence of outliers. In the presence of bivariate outliers, several robust alternatives exist for computing estimates of β_0 and β_1 . Two alternatives to OLS implemented in the MASS package will be considered. Specifically, least-trimmed squares using the function `lqs()` and robust regression using an M estimator with the function `r1m()` are discussed. Just as OLS sought to minimize the squared vertical distance between all of the Y_i s over all possible lines, least-trimmed squares minimizes the q smallest residuals over all possible lines where $q = \lfloor (n + p + 1)/2 \rfloor$. Fitting for the function `r1m()` is done by iterated re-weighted least squares. Although `lqs()` and `r1m()` are computationally intensive, the interfaces for `lm()`, `lqs()`, and `r1m()` are essentially identical. All three functions require a model formula of the form $y \sim x$. The \sim in this notation is read “is modeled by.”

Example 2.24 In Exercise 2.23 on the preceding page, the correlation between the logarithms of the body and brain weights in the data frame `Animals` from the MASS package with and without dinosaurs was computed. Find the estimates for the least squares regression lines with and without dinosaurs where the logarithm of brain is modeled by the logarithm of body using Equations (2.8) and (2.9) as well as the S function `lm()`. Superimpose both lines on the scatterplot using the function `abline()` (see Table A.12 on page 667).

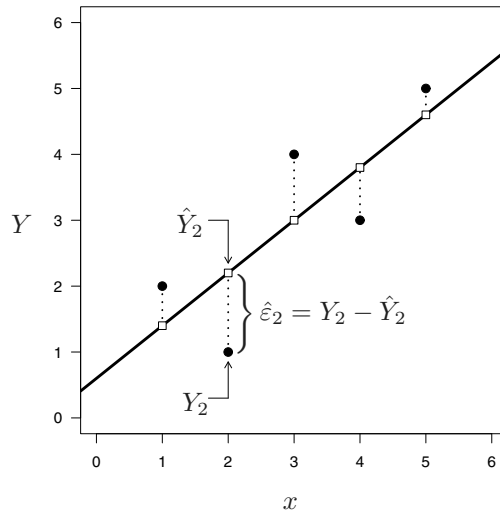


FIGURE 2.19: Graph depicting residuals. The vertical distances shown with a dotted line between the Y_i s, depicted with a solid circle, and the \hat{Y}_i s, depicted with a clear square, are the residuals.

Solution: Recall that there are a total of 28 animals in the data frame `Animals` and 25 animals in the `NoDINO` data frame. The scatterplot with superimposed regression lines including the dinosaurs and omitting the dinosaurs is shown in Figure 2.20 on the next page.

```
> attach(Animals)
> Y <- log(brain)
> X <- log(body)
> plot(X, Y, xlab="log(body)", ylab="log(brain)")
> b1 <- sum((X-mean(X))*(Y-mean(Y)))/sum((X-mean(X))^2)
> b0 <- mean(Y) - b1*mean(X)
> estimates <- c(b0, b1)
> estimates
[1] 2.5548981 0.4959947
> modDINO <- lm(Y~X)
> modDINO
```

Call:

```
lm(formula = Y ~ X)
```

Coefficients:

```
(Intercept)      X
      2.555      0.496
```

```
> abline(modDINO, col="pink", lwd=2)
> SA <- Animals[order(body),] # Sorted by body weight
> NoDINO <- SA[-(28:26),] # Remove rows 26-28 (dinosaurs)
```

```

> detach(Animals)
> attach(NoDINO)                # NoDINO contains 25 rows
> Y <- log(brain)
> X <- log(body)
> b1 <- sum((X-mean(X))*(Y-mean(Y)))/sum((X-mean(X))^2)
> b0 <- mean(Y) - b1*mean(X)
> estimates <- c(b0, b1)
> estimates
[1] 2.1504121 0.7522607
> modNODINO <- lm(Y~X)
> modNODINO

```

Call:

```
lm(formula = Y ~ X)
```

Coefficients:

(Intercept)	X
2.1504	0.7523

```

> abline(modNODINO, col="blue", lwd=2, lty=2)
> leglabels <- c("OLS with Dinosaurs", "OLS without Dinosaurs")
> leglty <- c(1,2)
> legcol=c("pink","blue")
> legend("topleft", legend=leglabels, lty=leglty, col=legcol, lwd=2)
> detach(NoDINO)

```

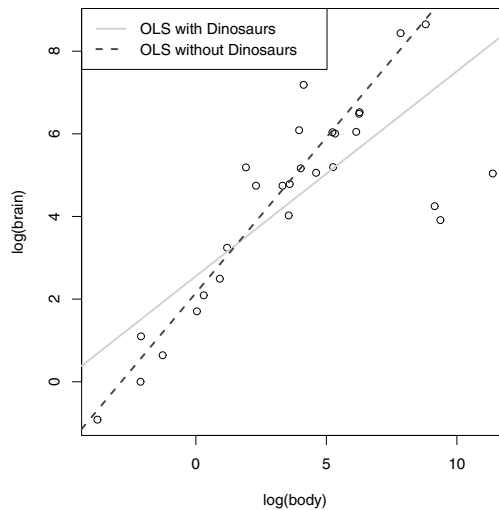


FIGURE 2.20: Scatterplot of $\log(\text{brain})$ versus $\log(\text{body})$ with superimposed regression lines computed with (solid line) and without (dashed line) dinosaurs

The intercept and slope of the regression line with dinosaurs are 2.555 and 0.496, respectively. Without the dinosaurs, the intercept and slope of the regression line are 2.1504 and 0.7523, respectively. ■

Example 2.25 From Figure 2.20 in Exercise 2.24 one notices three bivariate outliers (dinosaurs). Fit regression lines to the same data used in Exercise 2.20 using ordinary least squares, least-trimmed squares, and robust regression with an M estimator. Superimpose the resulting regression lines on a scatterplot and label the lines accordingly.

Solution: The scatterplot with the three superimposed regression lines is shown in Figure 2.21 on the facing page.

```
> attach(Animals)
> plot(log(body), log(brain), col="blue")
> f <- log(brain)~log(body)
> modelLM <- lm(f)
> modelLM
Call: lm(formula = f)

Coefficients:
(Intercept)    log(body)
      2.555         0.496

> abline(modelLM, col="pink", lwd=2)
> modelLQS <- lqs(f)
> modelLQS
Call: lqs.formula(formula = f)

Coefficients:
(Intercept)    log(body)
      1.816         0.776

Scale estimates 0.4637 0.4633

> abline(modelLQS, lty=2, col="red", lwd=2)
> modelRLM <- rlm(f, method="MM")
> modelRLM
Call: rlm(formula = f, method = "MM") Converged in 5 iterations

Coefficients:
(Intercept)    log(body)
  2.0486717    0.7512927

Degrees of freedom: 28 total; 26 residual
Scale estimate: 0.633

> abline(modelRLM, lty=3, col="black", lwd=2)
> leglabels <- c("Least Squares Line", "Least-Trimmed Squares",
+ "Robust line: M-estimator ")
> leglty <- c(1,2,3)
> legend("topleft", legend=leglabels, lty=leglty,
+ col=c("pink", "red", "black"), lwd=2, cex=0.85)
> detach(Animals)
```

The least-trimmed squares (`lqs()`) procedure and the robust line with M estimator (`rlm()`) method produce lines that put relatively little importance on outliers (dinosaurs). This is

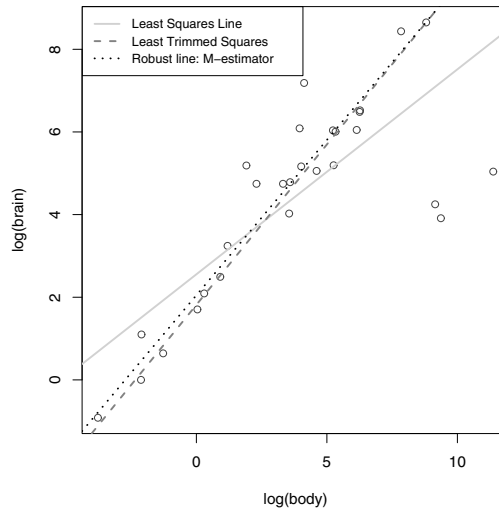


FIGURE 2.21: Scatterplot of $\log(\text{brain})$ versus $\log(\text{body})$ with three superimposed regression lines. Solid is the OLS line; dashed is the least-trimmed squares line; and dotted is the robust line.

further highlighted when one considers the estimates β_0 and β_1 for the OLS estimates without dinosaurs compared to the estimates of β_0 and β_1 for the least-trimmed squares and robust procedures given in Table 2.3.

Table 2.3: Different values for b_0 and b_1 with various regression methods

Method	b_0	b_1
OLS with dinosaurs	2.555	0.496
OLS without dinosaurs	2.150	0.752
least-trimmed squares	1.816	0.776
robust line with M estimator	2.049	0.751



2.8 Multivariate Data (Lattice and Trellis Graphs)

This section examines tools that can be used to understand multivariate data. Specifically, Trellis displays (used in S-PLUS), which were developed by Cleveland (1993), are introduced. The R version of Cleveland's Trellis displays is implemented with the package `lattice`. Trellis displays are graphs that examine higher dimensional structure in data by conditioning on one or more variables. Trellis graphs are implemented in a slightly different fashion from traditional S graphs; however, some readers may find the layout, rendering, and default coloring of Trellis graphs more appealing than traditional S graphs. Trellis

graphs are created with a formula syntax. The formula expresses the dependencies between the variables as follows:

$$\text{response} \sim \text{predictor} \mid \text{conditioning.variable}$$

The expression $y \sim x \mid z$ is read “ y is modeled as x given z .” Depending on the type of graph, all three components may not need to be specified. Table A.11 on page 666 lists the arguments for some of the more popular Trellis functions. If there is more than one conditioning variable, they are all listed separated by the multiplication symbol (*).

Example 2.26 Use Trellis histograms to compare the body weight index (BWI) for the two treatments (traditional sitting and hamstring stretch stored in `Treatment`) using the data frame `EPIDURAL`.

Solution: Recall that BWI is typically defined as kg/m^2 . Since the data frame `EPIDURAL` does not contain a BWI variable, one is created:

```
> attach(EPIDURAL)
> BWI <- kg/(cm/100)^2
> library(lattice) # only for R
> histogram(~BWI|Treatment, layout=c(1,2))
> detach(EPIDURAL)
```

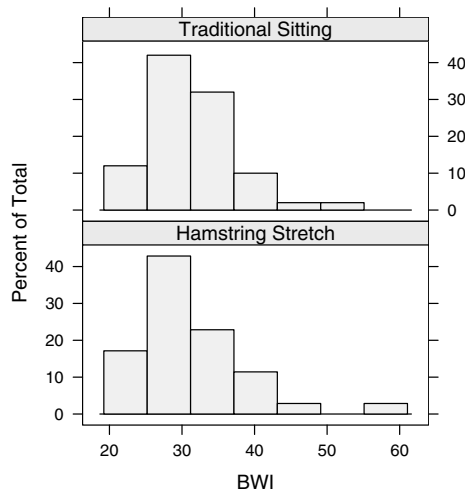


FIGURE 2.22: Comparative histograms of BWI by treatment

The `histogram()` function used the additional argument of `layout=c(1,2)`. The first value of `layout` determines the number of columns (1) in the Trellis graph and the second value determines the number of rows (2) in the Trellis graph. This is in contrast to how dimensions are specified in a matrix, which is number of rows by number of columns. The basic shapes of the two histograms shown in Figure 2.22 are quite similar, just as was observed in Example 2.18 on page 54 when the histograms were created using traditional S graphs. ■

Example 2.27 In Example 2.19 on page 54 side-by-side boxplots were used to compare the BWI for the two treatments. An additional concern is that not only should the distribution of BWI be similar for treatments, but it should also be similar for each physician. Use Trellis graphs to create side-by-side boxplots of BWI by treatments given Doctor using the data frame EPIDURAL.

Solution: The argument `as.table=TRUE` used in the `bwplot()` function orders the graphs the way one reads a book. The default arrangement of graphs is to start in the lower left and move to the upper right. This is done so that the graphs appear with the smallest values in the lower left, analogous to a scatterplot.

```
> attach(EPIDURAL)
> BWI <- kg/(cm/100)^2
> library(lattice)
> bwplot(Treatment~BWI|Doctor, as.table=TRUE) # Order: as one reads
```

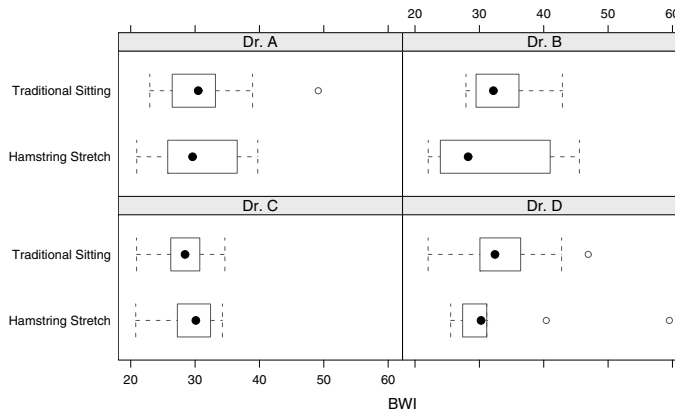


FIGURE 2.23: Trellis side-by-side boxplots of BWI in the traditional sitting and hamstring stretch positions given Doctor

Since the number of observations for each of the treatments is relatively small (range is from 6 to 15), it might be a better to look at the data with a **stripplot**. A stripplot of the treatments conditioning on physician is illustrated in Figure 2.24 on the following page.

```
> stripplot(Treatment~BWI|Doctor, jitter=TRUE, as.table=TRUE)
> detach(EPIDURAL)
```

The optional argument `jitter=TRUE` adds a small amount of noise to the values in the stripplot so that overlapping values are easier to distinguish. Based on the stripplots shown in Figure 2.24 on the next page, it seems that Dr. C's patients have a consistently smaller BWI for both treatment positions. Further investigation is needed to see why Dr. C's patients have consistently smaller BWI measurements versus the other physicians.

2.8.1 Arranging Several Graphs on a Single Page

The arrangement of Trellis graphs on a single page is again different from the arrangement of traditional graphs on a single page. Two different approaches can be taken when

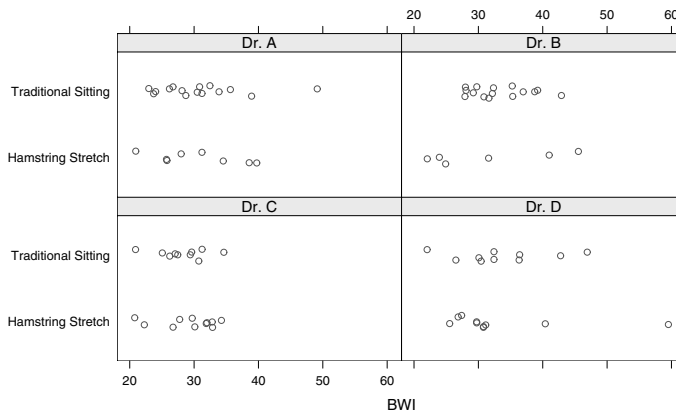


FIGURE 2.24: Trellis side-by-side stripplots of BWI in the traditional sitting and hamstring stretch positions given Doctor

arranging several graphs on a single page. The first approach discussed is to arrange the graphs in equally sized rectangles based on the dimensions of a matrix. In other words, if one wants to plot six graphs on a single page, it might be accomplished with a 3 by 2 or a 2 by 3 matrix where each position of the matrix represents a graph. To print each graph, the following structure is used:

```
print(trellisgraph, split=c(column, row, number of columns,
number of rows), more=TRUE/FALSE)
```

A second approach to producing multiple graphs on a single page is to literally specify the lower left and upper right coordinates for each graph. The lower left of the graph is denoted by the coordinates (0, 0), and the upper right corner is denoted by the coordinates (1, 1). The form for specifying each graph is $(x_{LL}, y_{LL}, x_{UR}, y_{UR})$. To print each graph, the following structure is used:

```
print(trellisgraph, position=c(x_LL, y_LL, x_UR, y_UR), more=TRUE/FALSE)
```

Example 2.28 Use Trellis graphs to create boxplots of BWI given Doctor, a scatterplot of cm versus kg given Doctor, a histogram of BWI, and a density plot of BWI given Treatment using the data frame EPIDURAL. Show all four graphs on the same page.

Solution: The solution provided is for R. The commands that follow will work in S-PLUS for graphs 2–4. However, the command `bwplot(~BWI|Doctor)` (graph 1) will not work in S-PLUS. The argument `as.table=TRUE` used in the `bwplot()` and the `xyplot()` functions are not requested in the problem. However, they are used since most people like to read from left to right and top to bottom. The four graphs are created and stored in variables named `graph1`, `graph2`, `graph3`, and `graph4`, respectively. By splitting the graph into a 2 by 2 matrix or by specifying the position for each of the four graphs one can reproduce Figure 2.25 on the facing page using the commands that follow.

```
> attach(EPIDURAL)
> library(lattice)
> graph1 <- bwplot(~BWI|Doctor, as.table=TRUE)
> graph2 <- xyplot(cm~kg|Doctor, as.table=TRUE)
```

```

> graph3 <- histogram(~BWI)
> graph4 <- densityplot(~BWI|Treatment)
> print(graph1, split=c(1,2,2,2), more=TRUE) # Lower left
> print(graph2, split=c(2,2,2,2), more=TRUE) # Lower right
> print(graph3, split=c(1,1,2,2), more=TRUE) # Upper left
> print(graph4, split=c(2,1,2,2), more=FALSE) # Upper right

```

Using the literal position of the graph

```

> print(graph1, position=c(0,0,.5,.5), more=TRUE) # Lower left
> print(graph2, position=c(.5,0,1,.5), more=TRUE) # Lower right
> print(graph3, position=c(0,.5,.5,1), more=TRUE) # Upper left
> print(graph4, position=c(.5,.5,1,1), more=FALSE) # Upper right
> detach(EPIDURAL)

```

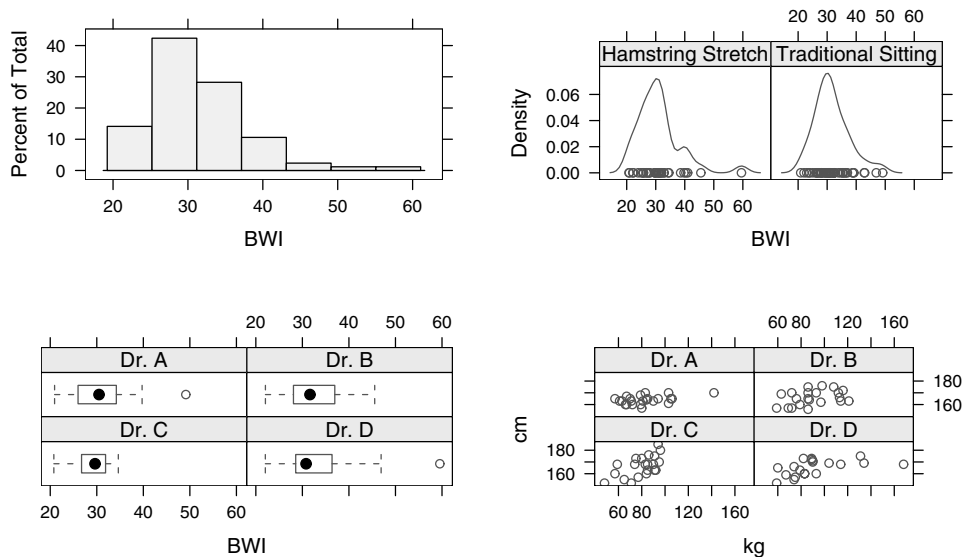


FIGURE 2.25: Arrangement of four different Trellis graphs on the same page

2.8.2 Panel Functions

Panel functions can be used to add additional features to a Trellis graph. For example, given a Trellis x - y plot, one can add a line using the panel function `panel.abline()`. For a list of available panel functions in R, type `?panel.functions` at the R prompt. For details with S-PLUS, see the S-PLUS Programmer's Guide.

Example 2.29 Create a Trellis x - y plot of `cm` versus `kg` given `Doctor` using the data frame `EPIDURAL`. Use panel functions to superimpose the ordinary least squares line and a least-trimmed squares line over the x - y plot.

Solution: The commands that follow create Figure 2.26.

```
> library(lattice)
> library(MASS) # Needed for lqs
> attach(EPIDURAL)
> xyplot(cm~kg|Doctor, as.table=TRUE,
+ panel=function(x, y)
+ {
+ panel.xyplot(x, y) # x-y plot
+ panel.abline(lm(y~x)) # Least sq line
+ panel.abline(lqs(y~x), col=3, lty=2, lwd=2) # Least trim sq line
+ }
+ )
```

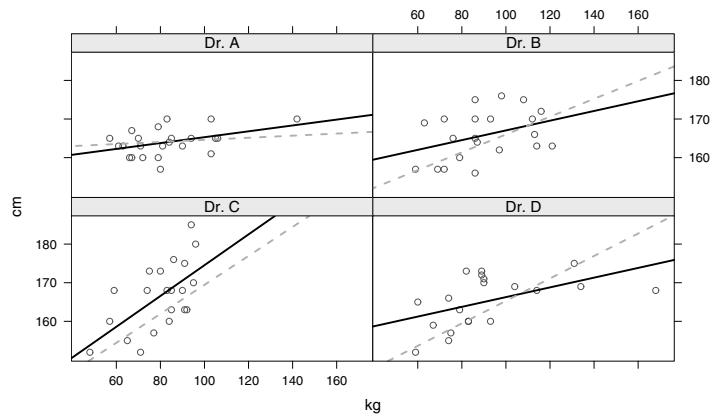


FIGURE 2.26: x - y plot of height (cm) versus weight (kg) given physician (Doctor) with superimposed least squares and least-trimmed squares lines

Another approach is to create a panel function that will superimpose the least squares and least-trimmed squares lines on an x - y plot and then to call that function within the `xyplot()` as follows:

```
> panel.scatregr <- function(x, y) # name function
+ {
+ panel.xyplot(x, y) # make x-y plot
+ panel.abline(lm(y~x), lwd=2) # regression line
+ panel.abline(lqs(y~x), col=3, lty=2, lwd=2) # least trim sq line
+ }
> xyplot(cm~kg|Doctor, as.table=TRUE, panel=panel.scatregr)
> detach(EPIDURAL)
```

Both approaches produce identical output. The dashed lines (`lty=2`) in Figure 2.26 are the least-trimmed squares lines. ■

2.9 Problems

1. Load the MASS package.
 - (a) Enter the command `help(package="MASS")` and read about the functions and data contained in this package.
 - (b) What does the description in the help file say about the function `lqs()`? Enter `help(lqs, package="MASS")` to obtain information about the command `lqs`.
 - (c) What command shows the loaded packages?
2. Load `Cars93` from the MASS package.
 - (a) Create density histograms for the variables `Min.Price`, `Max.Price`, `Weight`, and `Length` variables using a different color for each histogram.
 - (b) Superimpose estimated density curves over the histograms.
 - (c) Load the `lattice` package and do a box and whiskers plot of `Price` for every type of vehicle according to the drive train. Do you observe any differences between prices?
3. Load the data frame `WheatSpain` from the PASWR package.
 - (a) Find the quantiles, deciles, mean, maximum, minimum, interquartile range, variance, and standard deviation of the variable `hectares`. Comment on the results. What was Spain's 2004 total harvested wheat area in hectares?
 - (b) Create a function that calculates the quantiles, the mean, the variance, the standard deviation, the total, and the range of any variable.
 - (c) Which communities are below the 10th percentile in `hectares`? Which communities are above the 90th percentile? In which percentile is Navarra?
 - (d) Create and display in the same graphics device a frequency histogram of the variable `acres` and a density histogram of the variable `acres`. Superimpose a density curve over the second histogram.
 - (e) Explain why using breaks of 0; 100,000; 250,000; 360,000; and 1,550,000 automatically results in a density histogram.
 - (f) Create and display in the same graphics device a barplot of `acres` and a density histogram of `acres` using break points of 0; 100,000; 250,000; 360,000; and 1,550,000.
 - (g) Add vertical lines to the density histogram of `acres` to indicate the locations of the mean and the median, respectively.
 - (h) Create a boxplot of `hectares` and label the communities that appear as outliers in the boxplot. (Hint: Use `identify()`.)
 - (i) Determine the community with the largest harvested wheat surface area using either `acres` or `hectares`. Remove this community from the data frame and compute the mean, median, and standard deviation of `hectares`. How do these values compare to the values for these statistics computed in (a)?
4. Load the `wheatUSA2004` data frame from the PASWR package.

- (a) Find the quantiles, deciles, mean, maximum, minimum, interquartile range, variance, and standard deviation for the variable `ACRES`. Comment on what the most appropriate measures of center and spread would be for this variable. What is the USA's 2004 total harvested wheat surface area?
 - (b) Which states are below the 20th percentile? Which states are above the 80th percentile? In which quantile is WI (Wisconsin)?
 - (c) Create a frequency and a density histogram in the same graphics device using square plotting regions of the values in `ACRES`.
 - (d) Add vertical lines to the density histogram from (c) to indicate the location of the mean and the median.
 - (e) Create a boxplot of the `ACRES` and locate the outliers' communities and their values.
 - (f) Determine the state with the largest harvested wheat surface in acres. Remove this state from the data frame and compute the mean, median, and standard deviation of `ACRES`. How do these values compare to the values for these statistics computed in (a)?
5. The data frame `vit2005` in the `PASWR` package contains descriptive information and the appraised total price (in euros) for apartments in Vitoria, Spain.
- (a) Create a frequency table, a piechart, and a barplot showing the number of apartments grouped by the variable `out`. For you, which method conveys the information best?
 - (b) Characterize the distribution of the variable `totalprice`.
 - (c) Characterize the relationship between `totalprice` and `area`.
 - (d) Create a Trellis plot of `totalprice` versus `area` conditioning on `toilets`. Are there any outliers? Ignoring any outliers, between what two values of `area` do apartments have both one and two bathrooms?
 - (e) Use the `area` values reported in (d) to create a subset of apartments that have both one and two bathrooms. By how much does an additional bathroom increase the appraised value of an apartment? Would you be willing to pay for an additional bathroom if you lived in Vitoria, Spain?
6. Access the data from url
<http://www.stat.berkeley.edu/users/statlabs/data/babies.data>
 and store the information in an object named `BABIES` using the function `read.table()`. A description of the variables can be found at
<http://www.stat.berkeley.edu/users/statlabs/labs.html>.
- These data are a subset from a much larger study dealing with child health and development.
- (a) Create a "clean" data set that removes subjects if any observations on the subject are "unknown." Note that `bwt`, `gestation`, `parity`, `height`, `weight`, and `smoke` use values of 999, 999, 9, 99, 999, and 9, respectively, to denote "unknown." Store the modified data set in an object named `CLEAN`.
 - (b) Use the information in `CLEAN` to create a density histogram of the birth weights of babies whose mothers have never smoked (`smoke=0`) and another histogram placed directly below the first in the same graphics device for the birth weights of babies whose mothers currently smoke (`smoke=1`). Make the range of the x -axis 30 to 180 (ounces) for both histograms. Superimpose a density curve over each histogram.

- (c) Based on the histograms in (b), characterize the distribution of baby birth weight for both non-smoking and smoking mothers.
 - (d) What is the mean weight difference between babies of smokers and non-smokers? Can you think of any reasons not to use the mean as a measure of center to compare birth weights in this problem?
 - (e) Create side-by-side boxplots to compare the birth weights of babies whose mother's never smoked and those who currently smoke. Use traditional graphics (`boxplot()`) as well as Trellis/lattice graphs to create the boxplots (`bwplot()`).
 - (f) What is the median weight difference between babies who are firstborn and those who are not?
 - (g) Create a single graph of the densities for pre-pregnancy `weight` for mothers who have never smoked and for mothers who currently smoke. Make sure both densities appear on the same graphics device and place a color coded legend in the top right corner of the graph.
 - (h) Characterize the pre-pregnancy distribution of `weight` for mothers who have never smoked and for mothers who currently smoke.
 - (i) What is the mean pre-pregnancy weight difference between mothers who do not smoke and those who do? Can you think of any reasons not to use the mean as a measure of center to compare pre-pregnancy weights in this problem?
 - (j) Compute the body weight index (BWI) for each mother in `CLEAN`. Recall that BWI is defined as kg/m^2 (0.0254 m = 1 in., and 0.45359 kg = 1 lb.). Add the variables `weight` in kg, `height` in m, and BWI to `CLEAN` and store the result in `CLEANP`.
 - (k) Characterize the distribution of BWI.
 - (l) Group pregnant mothers according to their BWI quartile. Find the mean and standard deviation for baby birth weights in each quartile for mothers who have never smoked and those who currently smoke. Find the median and IQR for baby birth weights in each quartile for mothers who have never smoked and those who currently smoke. Based on your answers, would you characterize birth weight in each group as relatively symmetric or skewed? Create histograms and densities of `bwt` conditioned on BWI quartiles and whether the mother smokes to verify your previous assertions about the shape.
 - (m) Create side-by-side boxplots of `bwt` based on whether the mother smokes conditioned on BWI quartiles. Does this graph verify your findings in (l)?
 - (n) Does it appear that BWI is related to the birth weight of a baby? Create a scatterplot of birth weight (`bwt`) versus BWI while conditioning on BWI quartiles and whether the mother smokes to help answer the question.
 - (o) Replace baby birth weight (`bwt`) with gestation length (`gestation`) and answer questions (l), (m), and (n).
 - (p) Create a scatterplot of `bwt` versus `gestation` conditioned on BWI quartiles and whether the mother smokes. Fit straight lines to the data using `lm()`, `lqs()`, and `rlm()`; and display the lines in the scatterplots. What do you find interesting about the resulting graphs?
 - (q) Create a table of `smoke` by `parity`. Display the numerical results in a graph. What percent of mothers did not smoke during the pregnancy of their first child?
7. Some claim the final hours aboard the RMS Titanic were marked by class warfare; others claim it was characterized by male chivalry. The data frame `titanic3` from the `PASWR`

- package contains information pertaining to class status (`pclass`), survival of passengers (`survived`), and gender (`sex`), to name but a few. Based on the information in `titanic3`:
- Determine the fraction of survivors (`survived`) according to class (`pclass`).
 - Compute the fraction of survivors according to class and gender. Did men in first class or women in third class have a higher survival rate?
 - How would you characterize the distribution of `age`?
 - Were the median and mean ages for females who survived higher or lower than for females who did not survive? Report the median and mean ages as well as an appropriate measure of spread for each statistic.
 - Were the median and mean ages for males who survived higher or lower than for males who did not survive? Report the median and mean ages as well as an appropriate measure of spread for each statistic.
 - What was the age of the youngest female in first class who survived?
 - Do the data suggest class warfare, male chivalry, or some combination of both characterized the final hours aboard the Titanic? Feel free to explore other relationships based on the numbers in `titanic3` in answering this question.
8. Use the `Cars2004EU` data frame from the `PASWR` package which contains the numbers of cars per 1000 inhabitants (`cars`), the total number of known mortal accidents (`deaths`), and the country population/1000 (`population`) for the 25 member countries of the European Union for the year 2004.
- Compute the total number of cars per 1000 inhabitants in each country, and store the result in an object named `total.cars`. Determine the total number of known automobile fatalities in 2004 divided by the total number of cars for each country and store the result in an object named `death.rate`.
 - Create a barplot showing the automobile death rate for each of the European Union member countries. Make the bars increase in magnitude so that the countries with the smallest automobile death rates appear first.
 - Which country has the lowest automobile death rate? Which country has the highest automobile death rate?
 - Create a scatterplot of `population` versus `total.cars`. How would you characterize the relationship?
 - Find the least squares estimates for regressing `population` on `total.cars`. Superimpose the least squares line on the scatterplot from (d). What population does the least squares model predict for a country with a `total.cars` value of 19224.630? Find the difference between the population predicted from the least squares model and the actual population for the country with a `total.cars` value of 19224.630.
 - Create a scatterplot of `total.cars` versus `death.rate`. How would you characterize the relationship between the two variables?
 - Compute Spearman's rank correlation coefficient of `total.cars` and `death.rate`. (Hint: Use `cor(x, y, method="spearman")`.) What is this coefficient measuring?
 - Plot the logarithm of `total.cars` versus the logarithm of `death.rate`. How would you characterize the relationship?
 - What are the least squares estimates for the regression of `log(death.rate)` on `log(total.cars)`. Superimpose the least squares line on the scatterplot from

- (h). What death rate does the least squares model predict for a country with a `log(total.cars)` value of 9.863948? Make sure you express your answer in the same units as those used for `death.rate`.
9. The data frame `SurfaceSpain` in the PASWR package contains the surface area (km²) for seventeen autonomous Spanish communities.
- Use the function `merge()` to combine the data frames `WheatSpain` (from problem 3) and `SurfaceSpain` into a new data frame named `DataSpain`.
 - Create a variable named `surface.h` containing the surface area of each autonomous community in hectares. (Note: 100 hectares = 1 km².) Create a variable named `wheat.p` containing the percent surface area in each autonomous community dedicated to growing wheat. Add the newly created variables to the data frame `DataSpain` and store the result as a data frame with the name `DataSpain.m`.
 - Assign the names of the autonomous communities as row names for `DataSpain.m` and remove the variable `community` from the data frame.
 - Create a barplot showing the percent surface area dedicated to growing wheat for each of the seventeen Spanish autonomous communities. Arrange the communities by decreasing percentages.
 - Display the percent surface area dedicated to growing wheat for each of the seventeen Spanish autonomous communities using the function `dotchart()`. To read about `dotchart()`, type `?dotchart` at the command prompt. Do you prefer the barchart or the dotchart? Explain your answer.
 - Describe the relationship between the surface area in an autonomous community dedicated to growing wheat (`hectares`) and the total surface area of the autonomous community (`surface.h`).
 - Describe the relationship between the surface area in an autonomous community dedicated to growing wheat (`hectares`) and the percent of surface area dedicated to growing wheat out of the communities' total surface area (`wheat.p`).
 - Develop a model to predict the surface area in an autonomous community dedicated to growing wheat (`hectares`) based on the total surface area of the autonomous community (`surface.h`).

Chapter 3

General Probability and Random Variables

3.1 Introduction

One of the main objectives of statistics is to help make “good” decisions under conditions of uncertainty. Probability is one way to quantify outcomes that cannot be predicted with certainty. For example, when throwing two dice, the outcome of the experiment cannot be known before the dice are thrown. Random variables, as well as counting techniques, will facilitate the analysis of problems such as the example of throwing two dice. This chapter provides a brief introduction to counting techniques, axiomatic probability, random variables, and moment generating functions.

3.2 Counting Rules

One of the fundamental questions surrounding any experiment is how to know the number of possible ways the experiment may have taken place.

DEFINITION 3.1: Basic principle of counting — Suppose k experiments are to be performed such that the first can result in any one of n_1 outcomes; and if for each of these n_1 outcomes, there are n_2 possible outcomes of the second experiment; and if for each of the possible outcomes of the first two experiments, there are n_3 possible outcomes of the third experiment; and if ..., then there are $n_1 \times n_2 \times \cdots \times n_k$ possible outcomes for the k experiments.

Example 3.1 A computer store sells three brands of laptops. Each laptop is sold with a carrying case and four different options for upgrading RAM. Suppose the store only carries two styles of carrying cases. How many different combinations of laptop, carrying case, and RAM are possible?

Solution: According to the basic principle of counting, there are $3 \cdot 2 \cdot 4 = 24$ different combinations of laptop, carrying case, and RAM. ■

3.2.1 Sampling With Replacement

When working with finite samples, it is critical to distinguish between **sampling with replacement** and **sampling without replacement**. Sampling with replacement occurs when an object is selected and subsequently replaced before the next object is selected. Consequently, when sampling with replacement, the number of possible ordered samples of size m taken from a set of n objects is n^m .

Example 3.2 How many different license plates can be made from four digits?

Solution: First, note that there is no restriction forbidding repeated digits. That is, 0001, 0002, 0003, . . . , 9999 are all permissible. In essence, this translates to sampling with replacement. Since there are 10 choices for each of the four license plate digits, there are a total of $10 \times 10 \times 10 \times 10 = 10^4 = 10,000$ possible license plates. ■

3.2.2 Sampling Without Replacement

Sampling without replacement occurs when an object is not replaced after it has been selected. When sampling without replacement, the number of possible ordered samples of size m taken from a set of n objects is

$$P_{m,n} = n(n-1)(n-2) \cdots (n-m+1) = \frac{n!}{(n-m)!}.$$

Any ordered sequence of m objects taken from n distinct objects is called a **permutation** and is denoted $P_{m,n}$.

Example 3.3 How many different ways can the first three places be decided in a race with four runners?

Solution: The number of ways the first three places can be decided using the basic principle of counting is by reasoning as follows:

Any one of the four runners might arrive in first place (four outcomes for the first experiment). After the first runner crosses the finish line, there are three possible choices for second place (three outcomes for the second experiment). Then, after second place is decided, there are only two runners left (two outcomes for the third experiment). Consequently, there are $4 \cdot 3 \cdot 2 = 24$ possible ways to award the first three places. The problem may also be solved by applying the permutation formula:

$$P_{3,4} = \frac{4!}{(4-3)!} = \frac{4!}{1!} = 4 \cdot 3 \cdot 2 = 24. \quad \blacksquare$$

Example 3.4 How many ways can seven students form a line?

Solution: First, note that once a student is selected for a place in line, the number of students for subsequent orderings is diminished by one. That is, this is a problem where sampling is done without replacement. A useful strategy for this type of problem is actually to think through assigning the students to positions before using a formula (permutation in this case). If seven slots are drawn, then the reasoning is as follows:

There are seven ways a student can be assigned to the first slot. Once the first slot has been assigned, there are six possible ways to pick a student for the next slot. Continue with this logic until all of the students have been assigned a slot. Appealing to the basic principle of counting, it is seen that there are $7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 7! = 5040$ possible ways to form a line with seven students. This is the same number calculated by considering a permutation of seven things taken seven at a time $P_{7,7} = \frac{7!}{(7-7)!} = \frac{7!}{0!} = 5040$. Note that $0! = 1$. ■

When a subset of the n objects is indistinguishable, clearly the number of permutations is diminished. More to the point, when n objects have n_1 that are the same, n_2 that are the same, and so on until there are n_k that are the same, there are a total of

$$\frac{n!}{n_1! \cdot n_2! \cdots n_k!}$$

permutations of the n objects.

Example 3.5 How many different letter arrangements can be formed using the letters *DATA*?

Solution: Note that there are $4!$ permutations of the letters $D_1A_1T_1A_2$ when the two A 's are distinguished from each other. However, since the A 's are indistinguishable, there are only $\frac{4!}{2! \cdot 1! \cdot 1!} = 12$ possible permutations of the letters *DATA*. ■

3.2.3 Combinations

In many problems, selecting m objects from n total objects without regard to order is the scenario of interest. For example, when selecting a committee, the order of the committee is rarely important. That is, a committee consisting of John, Mary, and Paul is considered the same committee if the members are listed Mary, Paul, and John. An arrangement of m objects taken from n objects without regard to order is called a **combination**. The number of combinations of n distinct objects taken m at a time is denoted as $C_{m,n}$ or $\binom{n}{m}$ and is calculated as

$$C_{m,n} = \binom{n}{m} = \frac{n!}{m!(n-m)!}.$$

Example 3.6 A committee of three people is to be formed from a group of eight people. How many different committees are possible?

Solution: There are $C_{3,8} = \binom{8}{3} = \frac{8!}{3!(8-3)!} = 56$ possible committees. ■

Example 3.7 How many different three-letter sequences can be formed from the letters A, B, C, and D if

- letter repetition is not permissible and order matters?
- letter repetition is permissible and order matters?
- letter repetition is not permissible and sequences containing the same letters are considered equal regardless of letter order?

Solution: The answers are as follows:

- There are $P_{3,4} = 4 \cdot 3 \cdot 2 = 24$ possible sequences.
- Since letters may be used more than once, there are $4^3 = 64$ possible sequences.
- Since order does not matter, there are $C_{3,4} = \binom{4}{3} = 4$ possible sequences. ■

Example 3.8 If nine people are to be assigned into three committees of sizes two, three, and four, respectively, how many possible assignments are possible?

Solution: There are $\binom{9}{2}$ ways to pick the first committee. Once that committee is selected, there are seven members left from which a committee of size three is selected. So, there are $\binom{7}{3}$ ways to pick the second committee. Using the same logic, there are finally four members left from which one committee of size four must be selected. There is only one way to select the remaining committee, which is to select all of the remaining members to be on the committee. Using the basic rule of multiplication, there are a total of $\binom{9}{2} \times \binom{7}{3} \times \binom{4}{4} = 1260$ ways to form the three committees. To compute the final answer, the `S` commands `choose()`, `prod()`, or a combination of the two can be used.

```

> choose(9,2)*choose(7,3)*choose(4,4)
[1] 1260
> prod(9:1)/(prod(2:1)*prod(3:1)*prod(4:1))
[1] 1260
> choose(9,2)*(prod(7:1)/(prod(3:1)*prod(4:1)))
[1] 1260

```



3.3 Probability

3.3.1 Sample Space and Events

An **experiment** is any action or process that generates observations. The **sample space** of an experiment, denoted by Ω , is the set of all of the possible outcomes of an experiment. Although the outcome of an experiment cannot be known before it has taken place, it is possible to define the sample space for a given experiment. The sample space may be either finite or infinite. For example, the number of unoccupied seats in a train corresponds to a finite sample space. The number of passengers arriving at an airport also produces a finite sample space, assuming a one to one correspondence between arriving passengers and the natural numbers. The sample space for the lifetime of light bulbs, however, is infinite, since lifetime may be any positive value.

An **event** is any subset of the sample space, which is often denoted with the letter E . Events are said to be **simple** when they contain only one outcome; otherwise, events are considered to be **compound**. Consider an experiment where a single die is thrown. Since the die might show any one of six numbers, the sample space is written $\Omega = \{1, 2, 3, 4, 5, 6\}$; and any subset of Ω , such as $E_1 = \{\text{even numbers}\}$, $E_2 = \{2\}$, $E_3 = \{1, 2, 4\}$, $E_4 = \Omega$, or $E_5 = \emptyset$, is considered an event. Specifically, E_2 is considered a simple event while all of the remaining events are considered to be compound events. Event E_5 is known as the **empty set** or the **null set**, the event that does not contain any outcomes. In many problems, the events of interest will be formed through a combination of two or more events by taking **unions**, **intersections**, and **complements**.

3.3.2 Set Theory

The following definitions review some basic notions from set theory and some basic rules of probability that are not unlike the rules of algebra. For any two events E and F of a sample space Ω , define the new event $E \cup F$ (read E union F) to consist of all outcomes that are either in E or in F or in both E and F . In other words, the event $E \cup F$ will occur if either E or F occurs. In a similar fashion, for any two events E and F of a sample space Ω , define the new event $E \cap F$ (read E intersection F) to consist of all outcomes that are both in E and in F . Finally, the complement of an event E (written E^c) consists of all outcomes in Ω that are not contained in E .

Given events E, F, G, E_1, E_2, \dots , the commutative, associative, distributive, and DeMorgan's laws work as follows with the union and intersection operators:

1. Commutative laws

- for the union $E \cup F = F \cup E$
- for the intersection $E \cap F = F \cap E$

2. Associative laws

- for the union $(E \cup F) \cup G = E \cup (F \cup G)$
- for the intersection $(E \cap F) \cap G = E \cap (F \cap G)$

3. Distributive laws

- $(E \cap F) \cup G = (E \cup G) \cap (F \cup G)$
- $(E \cup F) \cap G = (E \cap G) \cup (F \cap G)$

4. DeMorgan's laws

- $\left(\bigcup_{i=1}^{\infty} E_i \right)^c = \bigcap_{i=1}^{\infty} E_i^c$
- $\left(\bigcap_{i=1}^{\infty} E_i \right)^c = \bigcup_{i=1}^{\infty} E_i^c$

3.3.3 Interpreting Probability

3.3.3.1 Relative Frequency Approach to Probability

Suppose an experiment can be performed n times under the same conditions with sample space Ω . Let $n(E)$ denote the number of times (in n experiments) that the event E occurs. The relative frequency approach to probability defines the probability of the event E , written $\mathbb{P}(E)$, as

$$\mathbb{P}(E) = \lim_{n \rightarrow \infty} \frac{n(E)}{n}.$$

Although the preceding definition of probability is intuitively appealing, it has a serious drawback. There is nothing in the definition to guarantee $\frac{n(E)}{n}$ converges to a single value. Instead of assuming $\frac{n(E)}{n}$ converges, which is a very complex assumption, the simpler and more self-evident axioms about probability commonly referred to as the **three axioms of probability** are used.

3.3.3.2 Axiomatic Approach to Probability

The Three Axioms of Probability

Consider an experiment with sample space Ω . For each event E of the sample space Ω , assume that a number $\mathbb{P}(E)$ is defined that satisfies the following three axioms:

1. $0 \leq \mathbb{P}(E) \leq 1$
2. $\mathbb{P}(\Omega) = 1$
3. For any sequence of mutually exclusive events E_1, E_2, \dots (that is $E_i \cap E_j = \emptyset$ for all $i \neq j$,

$$\mathbb{P} \left(\bigcup_{i=1}^{\infty} E_i \right) = \sum_{i=1}^{\infty} \mathbb{P}(E_i).$$

The following results are all easily derived using some combination of the three axioms of probability:

1. $\mathbb{P}(E^c) = 1 - \mathbb{P}(E)$

Proof: Note that E and E^c are always mutually exclusive. Since $E \cup E^c = \Omega$, by probability axioms 2 and 3, $1 = \mathbb{P}(\Omega) = \mathbb{P}(E \cup E^c) = \mathbb{P}(E) + \mathbb{P}(E^c)$.

2. $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F)$

Proof: Note that $E \cup F$ can be represented as the union of two mutually exclusive events, E and $(E^c \cap F)$. That is, $E \cup F = E \cup (E^c \cap F)$. Event F can also be represented as the union of two mutually exclusive events, $(E \cap F)$ and $(E^c \cap F)$. By probability axiom 3, $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(E^c \cap F)$ as well as $\mathbb{P}(F) = \mathbb{P}(E \cap F) + \mathbb{P}(E^c \cap F)$. By solving for $\mathbb{P}(E^c \cap F)$ in the second equation and substituting the answer into the first equation, the desired result of $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F)$ is obtained.

3. $\mathbb{P}(\emptyset) = 0$

Proof: Consider two events, E_1 and E_2 , where $E_1 = \Omega$ and $E_2 = \emptyset$. Note that $\Omega = E_1 \cup E_2$ and $E_1 \cap E_2 = \emptyset$. By probability axioms 2 and 3, $1 = \mathbb{P}(\Omega) = \mathbb{P}(E_1) + \mathbb{P}(E_2) = 1 + \mathbb{P}(\emptyset) \implies \mathbb{P}(\emptyset) = 0$.

4. If $E \subset F$, then $\mathbb{P}(E) \leq \mathbb{P}(F)$

Proof: Since $E \subset F$, it follows that $F = E \cup (E^c \cap F)$. Note that E and $(E^c \cap F)$ are mutually exclusive events. Consequently, appealing to probability axiom 3, $\mathbb{P}(F) = \mathbb{P}(E) + \mathbb{P}(E^c \cap F)$. Since $\mathbb{P}(E^c \cap F) \geq 0$ by probability axiom 1, it follows that $\mathbb{P}(F) \geq \mathbb{P}(E)$.

Example 3.9 \triangleright *Law of Complement: Birthday Problem* \triangleleft Suppose that a room contains m students. What is the probability that at least two of them have the same birthday? This is a famous problem with a counterintuitive answer. Assume that every day of the year is equally likely to be a birthday, and disregard leap years. That is, assume there are always $n = 365$ days to a year.

Solution: Let the event E denote two or more students with the same birthday. In this problem, it is easier to find E^c , as there are a number of ways that E can take place. There are a total of 365^m possible outcomes in the sample space. E^c can occur in $365 \times 364 \times \dots \times (365 - m + 1)$ ways. Consequently,

$$\mathbb{P}(E^c) = \frac{365 \times 364 \times \dots \times (365 - m + 1)}{365^m}$$

and

$$\mathbb{P}(E) = 1 - \frac{365 \times 364 \times \dots \times (365 - m + 1)}{365^m}.$$

The following S code can be used to create or modify a table such as Table 3.1 on the next page, which gives $\mathbb{P}(E)$ for $m = 10, 15, \dots, 50$:

```
> for (m in seq(10,50,5))
  print(c(m, 1 - prod(365:(365-m+1))/365^m))
```

Another approach that can be used to solve the problem is to enter

```
> m <- seq(10,50,5)
> P.E <- function(m){c(m,1-prod(365:(365-m+1))/365^m)}
> t(sapply(m, P.E))
```

■

Table 3.1: Probability of two or more students having the same birthday

m	$\mathbb{P}(E)$
10	0.1169482
15	0.2529013
20	0.4114384
25	0.5686997
30	0.7063162
35	0.8143832
40	0.8912318
45	0.9409759
50	0.9703736

Example 3.10 Given two events E and F , suppose that $\mathbb{P}(E) = 0.3$, $\mathbb{P}(F) = 0.5$, and $\mathbb{P}(E \cup F) = 0.6$. Find $\mathbb{P}(E \cap F)$.

Solution: Since $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F)$, $0.6 = 0.3 + 0.5 - \mathbb{P}(E \cap F)$. Thus, $\mathbb{P}(E \cap F) = 0.2$. ■

3.3.4 Conditional Probability

In this section, conditional probability is introduced, which is one of the more important concepts in probability theory. Quite often, one is interested in calculating probabilities when only partial information obtained from an experiment is available. In such situations, the desired probabilities are said to be conditional. Even when partial information is unavailable, often the desired probabilities can be computed using conditional probabilities. If E and F are any two events in a sample space Ω and $\mathbb{P}(E) \neq 0$, the **conditional probability** of F given E is defined as

$$\mathbb{P}(F|E) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(E)}. \quad (3.1)$$

It is left as an exercise for the reader to verify that $\mathbb{P}(F|E)$ satisfies the three axioms of probability.

Example 3.11 Suppose two fair dice are tossed where each of the 36 possible outcomes is equally likely to occur. Knowing that the first die shows a 4, what is the probability that the sum of the two dice equals 8?

Solution: The sample space for this experiment is given as $\Omega = \{(i, j), i = 1, 2, \dots, 6, j = 1, 2, \dots, 6\}$, where each pair (i, j) has a probability $1/36$ of occurring. Define “the sum of the dice equals 8” to be event F and “a 4 on the first toss” to be event E . Since $E \cap F$ corresponds to the outcome $(4, 4)$ with probability $\mathbb{P}(E \cap F) = 1/36$ and there are

six outcomes with a 4 on the first toss, $(4, 1), (4, 2), \dots, (4, 6)$, the probability of event E , $\mathbb{P}(E) = 6/36 = 1/6$ and the answer is calculated as

$$\mathbb{P}(F|E) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(E)} = \frac{1/36}{1/6} = \frac{1}{6}. \quad \blacksquare$$

Example 3.12 Suppose a box contains 50 defective light bulbs, 100 partially defective light bulbs (last only 3 hours), and 250 good light bulbs. If one of the bulbs from the box is used and it does not immediately go out, what is the probability the light bulb is actually a good light bulb?

Solution: The conditional probability the light bulb is good given that the light bulb is not defective is desired. Using (3.1), write

$$\mathbb{P}(\text{Good}|\text{Not Defective}) = \frac{\mathbb{P}(\text{Good})}{\mathbb{P}(\text{Not Defective})} = \frac{250/400}{350/400} = \frac{5}{7}. \quad \blacksquare$$

3.3.5 The Law of Total Probability and Bayes' Rule

An important tool for solving probability problems where the sample space can be considered a union of mutually exclusive events is the **Law of Total Probability**.

Law of Total Probability — Let F_1, F_2, \dots, F_n be such that $\bigcup_{i=1}^n F_i = \Omega$ and $F_i \cap F_j = \emptyset$ for all $i \neq j$, with $\mathbb{P}(F_i) > 0$ for all i . Then, for any event E ,

$$\mathbb{P}(E) = \sum_{i=1}^n \mathbb{P}(E \cap F_i) = \sum_{i=1}^n \mathbb{P}(E|F_i)\mathbb{P}(F_i). \quad (3.2)$$

At times, it is much easier to calculate the conditional probabilities $\mathbb{P}(E|F_i)$ for an appropriately selected F_i than it is to compute $\mathbb{P}(E)$ directly. When this happens, **Bayes' Rule** is used, which is derived using (3.1), to find the answer.

Bayes' Rule — Let F_1, F_2, \dots, F_n be such that $\bigcup_{i=1}^n F_i = \Omega$ and $F_i \cap F_j = \emptyset$ for all $i \neq j$, with $\mathbb{P}(F_i) > 0$ for all i . Then,

$$\mathbb{P}(F_j|E) = \frac{\mathbb{P}(E \cap F_j)}{\mathbb{P}(E)} = \frac{\mathbb{P}(E|F_j)\mathbb{P}(F_j)}{\sum_{i=1}^n \mathbb{P}(E|F_i)\mathbb{P}(F_i)}. \quad (3.3)$$

Example 3.13 \triangleright **Conditional Probability: Car Batteries** \triangleleft A car manufacturer purchases car batteries from two different suppliers. Supplier A provides 55% of the batteries and supplier B provides the rest. If 5% of all batteries from supplier A are defective and 4% of the batteries from supplier B are defective, determine the probability that a randomly selected battery is not defective. (See Figure 3.1 on the facing page.)

Solution: Let C correspond to the event “the battery does not work properly,” A to the event “the battery was supplied by A ,” and B to the event “the battery was supplied by B .” The Venn diagram in Figure 3.1 on the next page provides a graphical illustration of the sample space for this example. Since a working battery might come from either supplier A or B , A and B are disjoint events. Consequently, $\mathbb{P}(C) = \mathbb{P}(C \cap A) + \mathbb{P}(C \cap B)$. Given that

$$\begin{aligned} \mathbb{P}(A) &= 0.55, \mathbb{P}(C|A) = 0.05, & \mathbb{P}(C \cap A) &= \mathbb{P}(C|A)\mathbb{P}(A), \\ \mathbb{P}(B) &= 0.45, \mathbb{P}(C|B) = 0.04, & \text{and } \mathbb{P}(C \cap B) &= \mathbb{P}(C|B)\mathbb{P}(B), \end{aligned}$$

write $\mathbb{P}(C) = (0.05)(0.55) + (0.04)(0.45) = 0.0455$. Then, the probability that the battery works properly is $1 - \mathbb{P}(C) = 0.9545$. \blacksquare

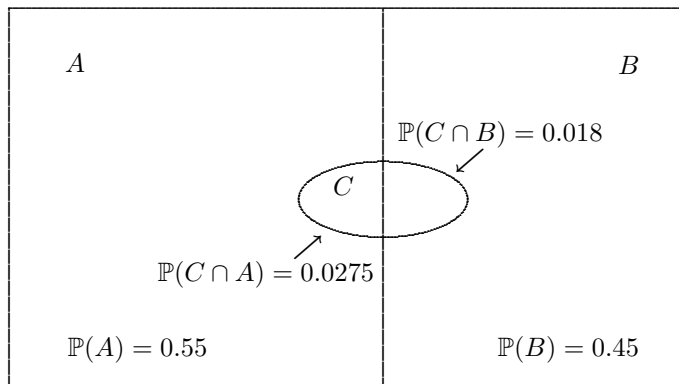


FIGURE 3.1: Sample space for Example 3.13

Example 3.14 Suppose a student answers all of the questions on a multiple-choice test. Let p be the probability the student actually knows the answer and $1 - p$ be the probability the student is guessing for a given question. Assume students that guess have a $1/a$ probability of getting the correct answer, where a represents the number of possible responses to the question. What is the conditional probability a student knew the answer to a question given that he answered correctly?

Solution: Let the events E , F_1 , and F_2 represent the events “question answered correctly,” “student knew the correct answer,” and “student guessed,” respectively. Using (3.3), write

$$\mathbb{P}(F_1|E) = \frac{\mathbb{P}(F_1 \cap E)}{\mathbb{P}(E)} = \frac{\mathbb{P}(F_1)}{\mathbb{P}(E|F_1)\mathbb{P}(F_1) + \mathbb{P}(E|F_2)\mathbb{P}(F_2)} = \frac{p}{p + (1 - p)/a}$$

As a special case, if $a = 4$ and $p = 1/2$, then the probability a student actually knew the answer given their response was correct is $4/5$. ■

Example 3.15 ▷ *Bayes’ Rule: Choose a Door* ◁ The television show *Let’s Make a Deal*, hosted by Monty Hall, gave contestants the chance to choose, among three doors, the one that concealed the grand prize. Behind the other two doors were much less valuable prizes. After the contestant chose one of the doors, say Door 1, Monty opened one of the other two doors, say Door 3, containing a much less valuable prize. The contestant was then asked whether he or she wished to stay with the original choice (Door 1) or switch to the other closed door (Door 2). What should the contestant do? Is it better to stay with the original choice or to switch to the other closed door? Or does it really matter? The answer, of course, depends on whether contestants improve their chances of winning by switching doors. In particular, what is the probability of winning by switching doors when given the opportunity; and what is the probability of winning by staying with the initial door selection? First, simulate the problem with S to provide approximate probabilities for the various strategies. Following the simulation, show how Bayes’ Rule can be used to solve the problem exactly.

Solution: To simulate the problem, generate a random vector named `actual` of size 10,000 containing the numbers 1, 2, and 3. In the vector `actual`, the numbers 1, 2, and 3 represent the door behind which the grand prize is contained. Then, generate another vector named `guess` of size 10,000 containing the numbers 1, 2, and 3 to represent the contestant’s initial guess. If the i^{th} values of the vectors `actual` and `guess` agree, the contestant wins the grand

prize by staying with his initial guess. On the other hand, if the i^{th} values of the vectors `actual` and `guess` disagree, the contestant wins the grand prize by switching. Consider the following S code and the results that suggest the contestant is twice as likely to win the grand prize by switching doors:

```
> actual <- sample(1:3, 10000, replace = TRUE)
> aguess <- sample(1:3, 10000, replace = TRUE)
> equals <- (actual == aguess)
> PNoSwitch <- sum(equals)/10000
> not.eq <- (actual != aguess)
> PSwitch <- sum(not.eq)/10000
> Probs <- c(PNoSwitch, PSwitch)
> names(Probs) <- c("P(Win no Switch)", "P(Win Switch)")
> Probs
      P(Win no Switch)  P(Win Switch)
                0.3317             0.6683
```

Next use (3.3) after defining events D_i and O_j to find $\mathbb{P}(D_1|O_3)$ and $\mathbb{P}(D_2|O_3)$. Start by assuming the contestant initially guesses Door 1 and that Monty opens Door 3. Let the event $D_i = \text{Door } i \text{ conceals the prize}$ and $O_j = \text{Monty opens door } j \text{ after the contestant selects Door 1}$. When a contestant initially selects a door, $\mathbb{P}(D_1) = \mathbb{P}(D_2) = \mathbb{P}(D_3) = 1/3$. Once Monty shows the grand prize is not behind Door 3, the probability of winning the grand prize is now one of $\mathbb{P}(D_1|O_3)$ or $\mathbb{P}(D_2|O_3)$. Note that $\mathbb{P}(D_1|O_3)$ corresponds to the strategy of sticking with the initial guess and $\mathbb{P}(D_2|O_3)$ corresponds to the strategy of switching doors. Based on how the show is designed, the following are known:

- $\mathbb{P}(O_3|D_1) = 1/2$ since Monty can open one of either Door 3 or Door 2.
- $\mathbb{P}(O_3|D_2) = 1$ since the only door Monty can open without revealing the grand prize is Door 3.
- $\mathbb{P}(O_3|D_3) = 0$ since Monty will not open Door 3 if it contains the grand prize.

$$\begin{aligned}\mathbb{P}(D_1|O_3) &= \frac{\mathbb{P}(O_3|D_1)\mathbb{P}(D_1)}{\mathbb{P}(O_3|D_1)\mathbb{P}(D_1) + \mathbb{P}(O_3|D_2)\mathbb{P}(D_2) + \mathbb{P}(O_3|D_3)\mathbb{P}(D_3)} \\ &= \frac{1/2 \times 1/3}{1/2 \times 1/3 + 1 \times 1/3 + 0 \times 1/3} = \frac{1}{3}\end{aligned}$$

$$\begin{aligned}\mathbb{P}(D_2|O_3) &= \frac{\mathbb{P}(O_3|D_2)\mathbb{P}(D_2)}{\mathbb{P}(O_3|D_1)\mathbb{P}(D_1) + \mathbb{P}(O_3|D_2)\mathbb{P}(D_2) + \mathbb{P}(O_3|D_3)\mathbb{P}(D_3)} \\ &= \frac{1 \times 1/3}{1/2 \times 1/3 + 1 \times 1/3 + 0 \times 1/3} = \frac{2}{3}\end{aligned}$$

Therefore, it is always to the contestant's benefit to switch doors. ■

3.3.6 Independent Events

Conditional probability allows for an alteration in the probability of an event when additional information is present. That is, $\mathbb{P}(E|F)$ is sometimes different from $\mathbb{P}(E)$ when some knowledge of the event F is available. Note that $\mathbb{P}(E|F)$ is *sometimes* different from $\mathbb{P}(E)$, not that it is *always* different. When $\mathbb{P}(E|F) = \mathbb{P}(E)$, clearly knowledge of the

event F does not alter the probability of obtaining E . When this happens, event E is **independent** of event F . More formally, two events E and F are independent if and only if $\mathbb{P}(E|F) = \mathbb{P}(E)$ or $\mathbb{P}(F|E) = \mathbb{P}(F)$. An equivalent way to define independence between two events is to use (3.1) and to show that $\mathbb{P}(E \cap F) = \mathbb{P}(E)\mathbb{P}(F)$. Independence between two events is really a special case of independence among n events. Define events E_1, \dots, E_n to be independent if, for every k where $k = 2, \dots, n$ and every subset of indices i_1, i_2, \dots, i_k , $\mathbb{P}(E_{i_1} \cap E_{i_2} \cap \dots \cap E_{i_k}) = \mathbb{P}(E_{i_1})\mathbb{P}(E_{i_2}) \dots \mathbb{P}(E_{i_k})$. It is important to point out that events in any subset of the original independent events of size r , where $r \leq k$, are also independent. Further, if events E_1, \dots, E_n are independent, then so are E_1^c, \dots, E_n^c .

Example 3.16 ▷ *Law of Probability: Components* ◁ A system consists of three components as illustrated in Figure 3.2. The entire system will work if either both components 1 and 2 work or if component 3 works. Components 1 and 2 are connected in series, while component 3 is connected in parallel with components 1 and 2. If all of the components function independently, and the probability each component works is 0.9, what is the probability the entire system functions?

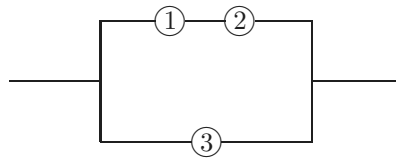


FIGURE 3.2: Circuit system diagram for Example 3.16

Solution: Let A_i ($i = 1, 2, 3$) be the event the i^{th} component works, and E the event the entire system works. Consequently, event $E = (A_1 \cap A_2) \cup A_3$, and $\mathbb{P}(E) = \mathbb{P}[(A_1 \cap A_2) \cup A_3]$.

$$\begin{aligned}
 \mathbb{P}(E) &= \mathbb{P}[(A_1 \cap A_2) \cup A_3] \\
 &= \mathbb{P}(A_1 \cap A_2) + \mathbb{P}(A_3) - \mathbb{P}(A_1 \cap A_2 \cap A_3) \\
 &= \mathbb{P}(A_1)\mathbb{P}(A_2) + \mathbb{P}(A_3) - \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3) \\
 &= (0.9)(0.9) + 0.9 - (0.9)(0.9)(0.9) \\
 &= 0.981
 \end{aligned}$$

■

3.4 Random Variables

In many experiments, it is easier to study some function of the outcomes than it is to study the original outcomes. For example, suppose 20 students are asked whether they favor legislation to reduce ozone emissions. Note that there are $2^{20} = 1,048,576$ possible outcomes in the sample space. However, it would make more sense to study the number of students who favor (equivalently, oppose) legislation out of 20 by defining a variable, say X , that equals the number of students favoring (or opposing) the legislation. Note that the sample space for X is the set of integers from 0 to 20, which is much easier to deal with than the original sample space. In general, a **random variable** is a function from a sample

space Ω into the real numbers. Random variables will always be denoted with uppercase letters, for example, X or Y , and the realized values of the random variable will be denoted with lowercase letters, for example, x or y . Here are some examples of random variables:

1. Toss two dice. X = the sum of the numbers on the dice.
2. A surgeon performs 20 heart transplants. X = the number of successful transplants.
3. Individual 40 kilometer cycling time trial. X = the time to complete the course.

Random variables may be either **discrete** or **continuous**. A random variable is said to be discrete if its set of possible outcomes is finite or at most countable. If the random variable can take on a continuum of values, it is continuous. Note that the random variables in examples 1 and 2 are discrete, while the variable in example 3 is continuous. If a random variable X has a distribution *DIST* with parameter(s) θ , write $X \sim \text{DIST}(\theta)$. If Y is a random variable that is distributed approximately *DIST* with parameter(s) θ , write $Y \sim \text{DIST}(\theta)$.

3.4.1 Discrete Random Variables

A discrete random variable assumes each of its values with a certain probability. When two dice are tossed, the probability the sum of two dice is 7, written $\mathbb{P}(X = 7)$, equals $1/6$. The function that assigns probability to the values of the random variable is called the probability density function, **pdf**. Many authors also refer to the **pdf** as the probability mass function (**pmf**) when working with discrete random variables. Denote the **pdf** as $p(x) = \mathbb{P}(X = x)$ for each x . All **pdfs** must satisfy the following two conditions:

1. $p(x) \geq 0$ for all x .
2. $\sum_{\forall x} p(x) = 1$.

The cumulative distribution function, **cdf**, is defined as

$$F(x) = \mathbb{P}(X \leq x) = \sum_{k \leq x} p(k).$$

Discrete **cdfs** have the following properties:

1. $0 \leq F(x) \leq 1$.
2. If $a < b$, then $F(a) \leq F(b)$ for any real numbers a and b . In other words, $F(x)$ is a non-decreasing function of x .
3. $\lim_{x \rightarrow \infty} F(x) = 1$.
4. $\lim_{x \rightarrow -\infty} F(x) = 0$.
5. $F(x)$ is a step function, and the height of the step at x is equal to $f(x) = \mathbb{P}(X = x)$.

Example 3.17 Toss a fair coin three times and let the random variable X represent the number of heads in the three tosses. Produce graphical representations of both the **pdf** and **cdf** for the random variable X .

Solution: The sample space for the experiment is

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

The random variable X can take on the values 0, 1, 2, and 3 with probabilities $\frac{1}{8}$, $\frac{3}{8}$, $\frac{3}{8}$, and $\frac{1}{8}$, respectively. Define the **cdf** for X , $F(x) = \mathbb{P}(X \leq x)$ as follows:

$$F(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1/8 & \text{if } 0 \leq x < 1 \\ 4/8 & \text{if } 1 \leq x < 2 \\ 7/8 & \text{if } 2 \leq x < 3 \\ 1 & \text{if } x \geq 3 \end{cases}$$

The code for producing a graph similar to Figure 3.3 on the next page with placement of specific values along the axes for both the **pdf** and **cdf** using the function `axis()` is as follows:

```
> x <- 0:3
> fx <- c(1/8,3/8,3/8,1/8)
> Fx <- c(1/8,4/8,7/8,1) # or Fx <- cumsum(fx)
> par(mfrow=c(1,2), pty="s")
> plot(x, fx, type="h", xlab="x", ylab="P(X=x)",
+ xlim=c(0,3), ylim=c(0,.4), xaxt="n", yaxt="n")
> axis(1, at=c(0,1,2,3), labels=c(0,1,2,3), las=1)
> axis(2, at=c(1/8,3/8), labels=c("1/8","3/8"), las=1)
> title("PDF")
> plot(x, Fx, type="n", xlab="x", ylab="F(x)",
+ xlim=c(-1,5), ylim=c(0,1), yaxt="n")
> axis(2, at=c(1/8,4/8,7/8,1), labels=c("1/8","4/8","7/8","1"), las=1)
> segments(-1,0,0,0)
> segments(0:4, c(Fx,1),1:5, c(Fx,1))
> lines(x, Fx, type="p", pch=16)
> segments(-1,1,5,1, lty=2)
> title("CDF")
```

■

3.4.2 Mode, Median, and Percentiles

The **mode** of a probability distribution is the x -value most likely to occur. If more than one such x value exists, the distribution is multimodal. The **median** of a distribution is the value m such that $\mathbb{P}(X \leq m) \geq 1/2$ and $\mathbb{P}(X \geq m) \geq 1/2$. The j^{th} **percentile** of a distribution is the value x_j such that $\mathbb{P}(X \leq x_j) \geq \frac{j}{100}$ and $\mathbb{P}(X \geq x_j) \geq 1 - \frac{j}{100}$. The m value that satisfies the definition for the median is not unique. If Example 3.17 on the facing page is considered, the modes are 1 and 2; and any value m between 1 and 2, not inclusive, satisfies the definition for the median. The 25th percentile of the distribution of X is 1 because $\mathbb{P}(X \leq 1) = \frac{4}{8} \geq \frac{25}{100}$ and $\mathbb{P}(X \geq 1) = \frac{7}{8} \geq 1 - \frac{25}{100}$.

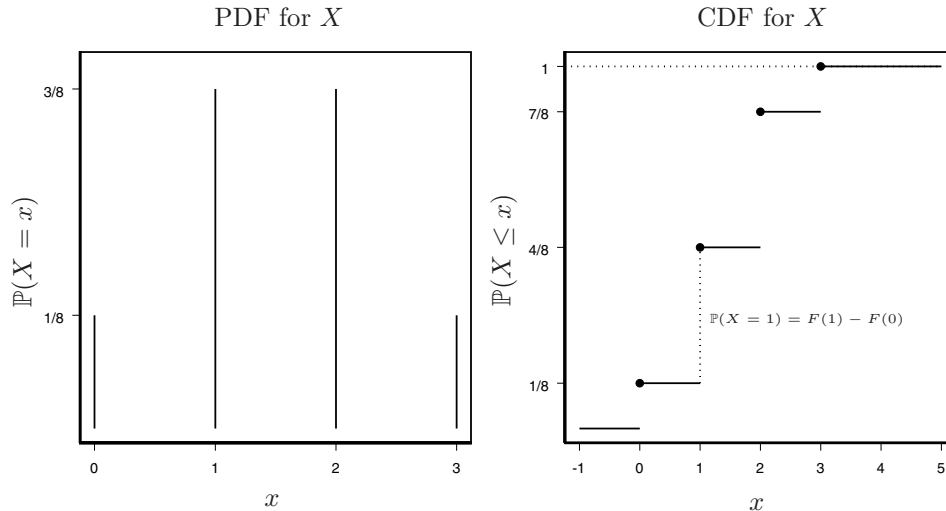


FIGURE 3.3: The **pdf** and **cdf** for the random variable X , the number of heads in three tosses of a fair coin

3.4.3 Expected Values of Discrete Random Variables

One of the more important ideas about summarizing the information provided in a **pdf** is that of expected value. Given a discrete random variable X with **pdf** $p(x)$, the **expected value** of the random variable X , written $E[X]$, is

$$E[X] = \sum_x x \cdot p(x) \quad (3.4)$$

Also denote $E[X]$ as μ_X , recognizing that $E[X]$ is the mean of the random variable X . In this definition, it is assumed the sum exists; otherwise, the expectation is undefined. It can be helpful to think of $E[X]$ as the fulcrum on a balance beam as illustrated in Figure 3.4.

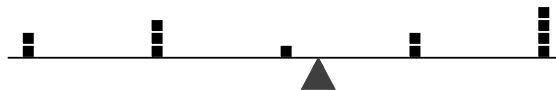


FIGURE 3.4: Fulcrum illustration of $E[X]$

Example 3.18 A particular game is played where the contestant spins a wheel that can land on the numbers 1, 5, or 30 with probabilities of 0.50, 0.45, and 0.05, respectively. The contestant pays \$5 to play the game and is awarded the amount of money indicated by the number where the spinner lands. Is this a fair game?

Solution: By fair, it is meant that the contestant should have an expected return equal to the price she pays to play the game. To answer the question, the expected (average)

winnings from playing the game need to be computed. Let the random variable X represent the player's winnings:

$$E[X] = \sum_x x \cdot p(x) = (1 \times 0.50) + (5 \times 0.45) + (30 \times 0.05) = 4.25.$$

Therefore, this game is not fair, as the house makes an average of 75 cents each time the game is played.

Another interpretation of the expected value of the random variable X is to view it as a weighted mean. Code to compute the expected value using (3.4) and using the function `weighted.mean()` is

```
> x <- c(1,5,30)
> px <- c(0.5,0.45,0.05)
> EX <- sum(x*px)
> WM <- weighted.mean(x, px)
> c(EX, WM)
[1] 4.25 4.25
```

Often, a random variable itself is not of interest, but rather some function of it is important, say $g(X)$, of the random variable X . The expected value of a function $g(X)$ of the random variable X with pdf $p(x)$ is

$$E[g(X)] = \sum_x g(x) \cdot p(x). \quad (3.5)$$

Example 3.19 Consider Example 3.18, for which the random variable Y is defined to be the player's net return. That is, $Y = X - 5$ since the player spends \$5 to play the game. What is the expected value of Y ?

Solution: The expected value of Y is

$$E[Y] = \sum_x (x - 5) \cdot p(x) = (-4 \times 0.50) + (0 \times 0.45) + (25 \times 0.05) = -0.75.$$

To compute the answer with S use

```
> x <- c(1,5,30)
> px <- c(0.5,0.45,0.05)
> EgX <- sum((x-5)*px)
> WgM <- weighted.mean((x-5), px)
> c(EgX, WgM)
[1] -0.75 -0.75
```

Rules of Expected Value The function $g(X)$ is often a linear function $a + bX$, where a and b are constants. When this occurs, $E[g(X)]$ is easily computed from $E[X]$. In Example 3.19, a and b were -5 and 1 , respectively, for the linear function $g(X)$. The following rules for expected value, when working with a random variable X and constants a and b , are true:

1. $E[bX] = bE[X]$.
2. $E[a + bX] = a + bE[X]$.

Unfortunately, if $g(X)$ is not a linear function of X , such as $g(X) = X^2$, the $E[X^2] \neq (E[X])^2$. In general, $E[g(X)] \neq g(E[X])$.

3.4.4 Moments

Another way to define the expected value of a random variable is with **moments**. However, knowing the mean (expected value) of a distribution does not tell the whole story. Several distributions may have the same mean. In this case, additional information, such as the spread of the distribution and the symmetry of the distribution, is helpful in distinguishing among various distributions.

The r^{th} **moment about the origin** of a random variable X , denoted α_r , is defined as $E[X^r]$. Note that $\alpha_1 = E[X^1]$ is called the mean of the distribution of X , also denoted μ_X or simply μ . The special moments defined next are important in the field of statistics as they help describe a random variable's distributional shape. The r^{th} **moment about the mean** of a random variable X , denoted μ_r , is the expected value of $(X - \mu)^r$. However, all moments do not exist. For the r^{th} moment about the origin of a discrete random variable to be well-defined, $\sum_{i=1}^{\infty} |x_i^r| \mathbb{P}(X = x_i)$ must be less than ∞ .

<p>Moments about 0 and μ</p> $E[X^r] = \alpha_r$ $E[(X - \mu)^r] = \mu_r$	(3.6)
--	-------

3.4.4.1 Variance

The second moment about the mean is called the **variance** of the distribution of X , or simply the variance of X :

$$\text{Var}[X] = \sigma_X^2 = E[(X - \mu)^2] = E[X^2] - \mu^2 \quad (3.7)$$

The positive square root of the variance is called the **standard deviation** and is denoted σ_X . The units of measurement for standard deviation are always the same as those for the random variable X . One way to avoid this unit dependency is to use the **coefficient of variation**, a unitless measure of variability.

DEFINITION 3.2: Coefficient of variation — When $E[X] \neq 0$,

$$CV_X = \frac{\sigma_X}{|E[X]|}, \quad (3.8)$$

3.4.4.2 Rules of Variance

If X is a random variable with mean μ and a and b are constants, then

1. $\text{Var}[b] = 0$.
2. $\text{Var}[aX] = a^2 \text{Var}[X]$.
3. $\text{Var}[aX + b] = a^2 \text{Var}[X]$.

Note that once $\text{Var}[aX + b] = a^2 \text{Var}[X]$ is proved, $\text{Var}[b] = 0$ and $\text{Var}[aX] = a^2 \text{Var}[X]$ have been implicitly shown.

Proof:

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

3.4.5 Continuous Random Variables

Recall that discrete random variables can only assume a countable number of outcomes. When a random variable has a set of possible values that is an entire interval of numbers, X is a **continuous random variable**. For example, if a 12 ounce can of beer is randomly selected and its actual fluid contents X is measured, then X is a continuous random variable because any value for X between 0 and the capacity of the beer can is possible.

Continuous Probability Density Functions' Properties

The function $f(x)$ is a **pdf** for the continuous random variable X , defined over the set of real numbers \mathbb{R} , if

1. $f(x) \geq 0, -\infty < x < \infty,$
2. $\int_{-\infty}^{\infty} f(x) dx = 1,$ and
3. $\mathbb{P}(a \leq X \leq b) = \int_a^b f(x) dx.$

(3.9)

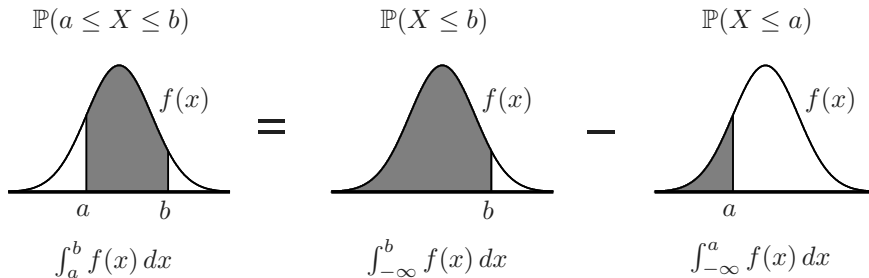


FIGURE 3.5: Illustration of $\mathbb{P}(a \leq X \leq b) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a)$

Condition 3 from (3.9) for the definition of a **pdf** for a continuous random variable is illustrated in Figure 3.5.

DEFINITION 3.3: Cumulative Density Function — The **cdf**, $F(x)$, of a continuous random variable X with **pdf** $f(x)$ is

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x f(t) dt, \quad -\infty < x < \infty \tag{3.10}$$

According to Definition 3.3, the **cdf** is derived from an existing **pdf**. Further, according

to the fundamental theorem of calculus, the other direction is also true since $F'(x) = f(x)$ for all values of x for which the derivative $F'(x)$ exists.

Continuous Cumulative Distribution Functions' Properties

Continuous **cdfs** have the following properties:

1. $0 \leq F(x) \leq 1$.
 2. If $a < b$, then $F(a) \leq F(b)$ for any real numbers a and b . In other words, $F(x)$ is a non-decreasing function of x .
 3. $\lim_{x \rightarrow \infty} F(x) = 1$.
 4. $\lim_{x \rightarrow -\infty} F(x) = 0$.
- (3.11)

Example 3.20 ▷ *Calculations of pdf and cdf* ◁ Suppose X is a continuous random variable with **pdf** $f(x)$, where

$$f(x) = \begin{cases} k(1 - x^2) & \text{if } -1 < x \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

- (a) Find the constant k so that $f(x)$ is a **pdf** of the random variable X .
- (b) Find the **cdf** for X .
- (c) Compute $\mathbb{P}(-0.5 \leq X \leq 1)$.
- (d) Graph the **pdf** and **cdf** of X with S.

Solution: The answers are as follows:

(a) Using property 2 from (3.9) for the **pdf** of a continuous random variable, write

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} f(x) dx = \int_{-1}^1 k(1 - x^2) dx \\ &= k \left[x - \frac{x^3}{3} \right]_{-1}^1 = k \left[\left(1 - \frac{1}{3} \right) - \left(-1 - \frac{-1}{3} \right) \right] \\ &= k \left[\frac{2}{3} - \frac{-2}{3} \right] = k \frac{4}{3} \Rightarrow k = \frac{3}{4}. \end{aligned}$$

(b) Using (3.3) it is known that

$$F(x) = \begin{cases} 0 & \text{if } x \leq -1 \\ \int_{-1}^x \frac{3}{4}(1 - t^2) dt = \frac{-x^3}{4} + \frac{3x}{4} + \frac{1}{2} & \text{if } -1 < x \leq 1 \\ 1 & \text{if } x > 1 \end{cases}$$

(c) Using property 3 from (3.9) for the **pdf** of a continuous random variable, write

$$\begin{aligned}
 \mathbb{P}(-0.5 \leq X \leq 1) &= F(1) - F(-0.5) \\
 &= \left(\frac{-1^3}{4} + \frac{3 \cdot 1}{4} + \frac{1}{2} \right) - \left(-\frac{\left(\frac{-1}{2}\right)^3}{4} + \frac{3 \cdot \frac{-1}{2}}{4} + \frac{1}{2} \right) \\
 &= \left(\frac{-1}{4} + \frac{3}{4} + \frac{1}{2} \right) - \left(\frac{1}{32} + \frac{-3}{8} + \frac{1}{2} \right) \\
 &= 1 - \frac{5}{32} = \frac{27}{32} = 0.84375.
 \end{aligned}$$

(d) Figure 3.6 depicts the **pdf** and **cdf** of X .

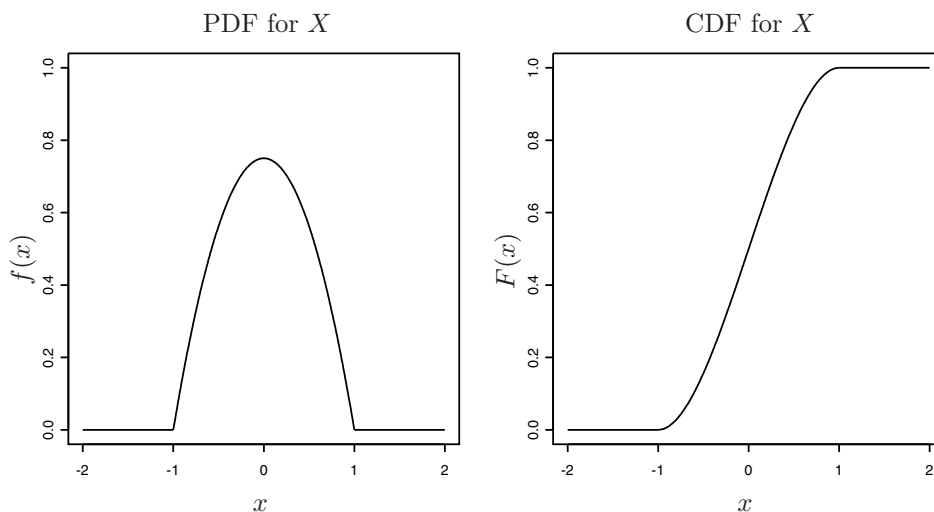


FIGURE 3.6: Illustration of **pdf** and **cdf** for Example 3.20

The following S code is used to create Figure 3.6:

```

> par(mfrow=c(1,2), pty="s")
> x <- seq(-1,1,0.01)
> y <- 3/4*(1-x^2)
> plot(x, y, xlim=c(-2,2), ylim=c(0,1), type="l", xlab="x",
+ ylab="f(x)")
> segments(-2,0,-1,0)
> segments(1,0,2,0)
> title("PDF for X")
> y <- -x^3/4 + 3*x/4 + 1/2
> plot(x, y, xlim=c(-2,2), ylim=c(0,1), type="l", xlab="x",
+ ylab="F(x)")
> segments(-2,0,-1,0)
> segments(1,1,2,1)
> title("CDF for X")

```



3.4.5.1 Numerical Integration with S

The S function `integrate()` approximates the integral of functions of one variable over a finite or infinite interval and estimates the absolute error in the approximation. To use `integrate()`, the user must specify `f()`, the function; `lower`, the lower limit of integration; and `upper`, the upper limit of integration. The function `f()` must be a real-valued S function of the form $f(x)$, where x is the variable of integration. In addition to using property 3 from (3.9) for the **pdf** of a continuous random variable to solve (c) of Example 3.20 on page 94, the problem could be solved directly by integrating the original probability $\mathbb{P}(-0.5 \leq X \leq 1)$. That is,

$$\mathbb{P}(-0.5 \leq X \leq 1) = \int_{-0.5}^1 \frac{3}{4} (1 - x^2) dx = \frac{3x}{4} - \frac{x^3}{4} \Big|_{-0.5}^1 = 0.84375.$$

The following code computes $\mathbb{P}(-0.5 \leq X \leq 1)$ using the function `integrate()` for R and S-PLUS, respectively:

```
> fx <- function(x){3/4-3/4*x^2}
> integrate(fx, lower=-0.5, upper=1)           # R
0.84375 with absolute error < 9.4e-15

> fx <- function(x){3/4-3/4*x^2}
> integrate(fx, lower=-0.5, upper=1)$integral # S-PLUS
[1] 0.84375
```

3.4.5.2 Mode, Median, and Percentiles

The **mode** of a continuous probability distribution, just like the mode of a discrete probability distribution, is the x -value most likely to occur. If more than one such x value exists, the distribution is multimodal. The **median** of a continuous distribution is the value m such that

$$\int_{-\infty}^m f(x) dx = \int_m^{\infty} f(x) dx = \frac{1}{2}.$$

The j^{th} **percentile** of a continuous distribution is the value x_j such that

$$\int_{-\infty}^{x_j} f(x) dx = \frac{j}{100}.$$

Example 3.21 Given a random variable X with **pdf**

$$f(x) = \begin{cases} 2e^{-2x} & \text{if } x > 0 \\ 0 & \text{if } x \leq 0, \end{cases}$$

- Find the median of the distribution.
- Find the 25th percentile of the distribution.
- Find the 60th percentile of the distribution.

Solution: The answers are as follows:

(a) The median is the value m such that $\int_0^m 2e^{-2x} dx = 0.5$, which implies

$$\begin{aligned} -e^{-2x} \Big|_0^m &= 0.5 \\ -e^{-2m} + 1 &= 0.5 \\ -e^{-2m} &= 0.5 - 1 \\ \ln(e^{-2m}) &= \ln(0.5) \\ m &= \frac{\ln(0.5)}{-2} = 0.3466. \end{aligned}$$

(b) The 25th percentile is the value x_{25} such that $\int_0^{x_{25}} 2e^{-2x} dx = 0.25$, which implies

$$\begin{aligned} -e^{-2x} \Big|_0^{x_{25}} &= 0.25 \\ -e^{-2x_{25}} + 1 &= 0.25 \\ -e^{-2x_{25}} &= 0.25 - 1 \\ \ln(e^{-2x_{25}}) &= \ln(0.75) \\ x_{25} &= \frac{\ln(0.75)}{-2} = 0.1438. \end{aligned}$$

(c) The 60th percentile is the value x_{60} such that $\int_0^{x_{60}} 2e^{-2x} dx = 0.60$, which implies

$$\begin{aligned} -e^{-2x} \Big|_0^{x_{60}} &= 0.60 \\ -e^{-2x_{60}} + 1 &= 0.60 \\ -e^{-2x_{60}} &= 0.60 - 1 \\ \ln(e^{-2x_{60}}) &= \ln(0.40) \\ x_{60} &= \frac{\ln(0.40)}{-2} = 0.4581. \end{aligned}$$

Example 3.22 Given a random variable X with pdf

$$f(x) = \begin{cases} 2 \cos(2x) & \text{if } 0 < x < \pi/4 \\ 0 & \text{otherwise,} \end{cases}$$

- Find the mode of the distribution.
- Find the median of the distribution.
- Draw the pdf and add vertical lines to indicate the values found in part (b).

Solution: The answers are as follows:

(a) The function $2 \cos 2x$ does not have a maximum in the open interval $(0, \pi/4)$ since the derivative $f'(x) = -4 \sin 2x$ does not equal 0 in the open interval $(0, \pi/4)$.

(b) The median is the value m such that

$$\int_0^m 2 \cos 2x \, dx = 0.5$$

$$\Downarrow$$

$$\sin 2x \Big|_0^m = \sin 2m = 0.5$$

$$2m = \arcsin(0.5)$$

$$m = \frac{\pi}{12}$$

(c) The R commands used to create Figure 3.7 are

```
> curve(2*cos(2*x),0, pi/4)
> abline(v=pi/12, lty=2, lwd=2)
```

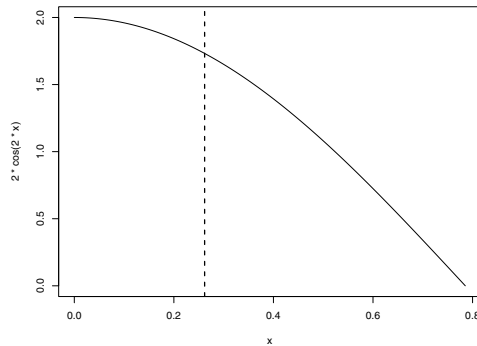


FIGURE 3.7: Graph of $2 \cos(2x)$ from 0 to $\frac{\pi}{4}$ with R

■

3.4.5.3 Expectation of Continuous Random Variables

For continuous random variables, the definitions associated with the expectation of a random variable X or a function, say $g(X)$, of X are identical to those for discrete random variables, except the summations are replaced with integrals and the probability density functions are represented with $f(x)$ instead of $p(x)$. The **expected value** of a continuous random variable X is

$$E[X] = \mu_X = \int_{-\infty}^{\infty} x \cdot f(x) \, dx. \quad (3.12)$$

When the integral in (3.12) does not exist, neither does the expectation of the random variable X . The expected value of a function of X , say $g(X)$, is

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f(x) \, dx. \quad (3.13)$$

Using the definitions for moments about 0 and μ given in (3.6), which relied strictly on expectation in conjunction with (3.13), the **variance** of a continuous random variable X is written as

$$\text{Var}[X] = \sigma_X^2 = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx. \quad (3.14)$$

Example 3.23 Given the function

$$f(x) = k, \quad -1 < x < 1$$

of the random variable X ,

- (a) Find the value of k to make $f(x)$ a **pdf**. Use this k for parts (b) and (c).
- (b) Find the mean of the distribution using (3.12).
- (c) Find the variance of the distribution using (3.14).

Solution: The answers are as follows:

- (a) Since $\int_{-\infty}^{\infty} f(x) dx$ must equal 1 for $f(x)$ to be a **pdf**, set $\int_{-1}^1 k dx$ equal to one and solve for k :

$$\begin{aligned} \int_{-1}^1 k dx &= 1 \\ kx \Big|_{-1}^1 &= 1 \\ 2k &= 1 \Rightarrow k = \frac{1}{2}. \end{aligned}$$

- (b) The mean of the distribution using (3.12) is

$$\begin{aligned} E[X] = \mu_X &= \int_{-1}^1 \frac{1}{2} x dx \\ &= \frac{x^2}{4} \Big|_{-1}^1 = 0 \end{aligned}$$

- (c) The variance of the distribution using (3.14) is

$$\begin{aligned} \text{Var}[X] = \sigma_X^2 &= E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx \\ &= \int_{-1}^1 (x - 0)^2 \frac{1}{2} dx \\ &= \frac{x^3}{6} \Big|_{-1}^1 = \frac{1}{3} \end{aligned}$$

■

3.4.6 Markov's Theorem and Chebyshev's Inequality

Theorem 3.1 Markov's Theorem If X is a random variable and $g(X)$ is a function of X such that $g(X) \geq 0$, then, for any positive K ,

$$\mathbb{P}(g(X) \geq K) \leq \frac{E[g(X)]}{K}. \quad (3.15)$$

Proof:

Step 1. Let $I(g(X))$ be a function such that

$$I(g(X)) = \begin{cases} 1 & \text{if } g(X) \geq K, \\ 0 & \text{otherwise.} \end{cases}$$

Step 2. Since $g(X) \geq 0$ and $I(g(X)) \leq 1$, when the first condition of $I(g(X))$ is divided by K ,

$$I(g(X)) \leq \frac{g(X)}{K}.$$

Step 3. Taking the expected value,

$$E[I(g(X))] \leq \frac{E[g(X)]}{K}.$$

Step 4. Clearly

$$\begin{aligned} E[I(g(X))] &= \sum_x I(g(x)) \cdot p(x) \\ &= [1 \cdot \mathbb{P}(I(g(X)) = 1)] + [0 \cdot \mathbb{P}(I(g(X)) = 0)] \\ &= [1 \cdot \mathbb{P}(g(X) \geq K)] + [0 \cdot \mathbb{P}(g(X) < K)] \\ &= \mathbb{P}(g(X) \geq K). \end{aligned}$$

Step 5. Rewriting,

$$\mathbb{P}(g(X) \geq K) \leq \frac{E[g(X)]}{K},$$

which is the inequality from (3.15) to be proven.

If $g(X) = (X - \mu)^2$ and $K = k^2\sigma^2$ in (3.15), it follows that

$$\mathbb{P}((X - \mu)^2 \geq k^2\sigma^2) \leq \frac{E[(X - \mu)^2]}{k^2\sigma^2} = \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}. \quad (3.16)$$

Working inside the probability on the left side of the inequality in (3.16), note that

$$\begin{aligned} ((X - \mu)^2 \geq k^2\sigma^2) &\Rightarrow (X - \mu \geq \sqrt{k^2\sigma^2}) \text{ or } (X - \mu \leq -\sqrt{k^2\sigma^2}) \\ &\Rightarrow (|X - \mu| \geq \sqrt{k^2\sigma^2}) \\ &\Rightarrow (|X - \mu| \geq k\sigma). \end{aligned}$$

Using this, rewrite (3.16) to obtain

$$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}, \quad (3.17)$$

which is known as **Chebyshev's Inequality**.

DEFINITION 3.4: Chebyshev's Inequality — Can be stated as any of

$$(a) \mathbb{P}(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}.$$

$$(b) \mathbb{P}(|X - \mu| < k) \geq 1 - \frac{\sigma^2}{k^2}.$$

$$(c) \mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

$$(d) \mathbb{P}(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}.$$

Version (d) of Chebyshev's Inequality is the complement of (c), the version derived in (3.17). Version (b) is the complement of (a) both of which can be obtained by setting $g(X) = (X - \mu)^2$ and $K = k^2$ in (3.15). A verbal interpretation of version (c) is that the probability any random variable X with finite variance, irrespective of the distribution of X , is k or more standard deviations from its mean is less than or equal to $1/k^2$. Likewise, version (d) states that the probability X is within k standard deviations from the mean is at least $1 - \frac{1}{k^2}$. Clearly, Chebyshev's Inequality can be used as a bound for certain probabilities. However, in many instances, the bounds provided by the inequality are very conservative. One reason for this is that there are no restrictions on the underlying distribution.

Example 3.24 Consider Example 3.17 on page 88, where X was defined to be the number of heads in three tosses of a fair coin. Chebyshev's Inequality guarantees at least what fraction of the distribution of X is within $k = 2$ standard deviations from its mean? What is the actual fraction of the distribution of X that is within $k = 2$ standard deviations from its mean?

Solution: Using version (d) of Chebyshev's Inequality, $\mathbb{P}(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}$, compute the first answer to be $1 - \frac{1}{2^2} = \frac{3}{4}$. To answer the second question, first find the mean and variance of X :

$$\begin{aligned} E[X] &= \sum_x x p(x) = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = \frac{3}{2} = 1.5 \\ E[X^2] &= \sum_x x^2 p(x) = 0^2 \times \frac{1}{8} + 1^2 \times \frac{3}{8} + 2^2 \times \frac{3}{8} + 3^2 \times \frac{1}{8} = \frac{3}{2} = 3 \\ \text{Var}[X] &= E[X^2] - (E[X])^2 = 3 - 1.5^2 = 0.75 \end{aligned}$$

For this example,

$$\begin{aligned} \mathbb{P}(|X - \mu| < k\sigma) &= \mathbb{P}(|X - 1.5| < 2\sqrt{0.75}) \\ &= \mathbb{P}(|X - 1.5| < 1.732) \\ &= \mathbb{P}(-0.232 < X < 3.232) = 1. \end{aligned}$$

Chebyshev's Inequality guaranteed at least 75% of the distribution of X would be within $k = 2$ standard deviations from its mean. However, the fact that all of the distribution of

X is within $k = 2$ standard deviations from the mean illustrates the conservative nature of Chebyshev's Inequality.

To compute the needed quantities with S, use the following code:

```
> x <- 0:3
> px <- c(1/8,3/8,3/8,1/8)
> EX <- weighted.mean(x, px)
> EX2 <- weighted.mean(x^2, px)
> VX <- EX2 - EX^2
> sigmaX <- sqrt(VX)
> c(EX, EX2, VX, sigmaX)
[1] 1.5000000 3.0000000 0.7500000 0.8660254
```

■

3.4.7 Weak Law of Large Numbers

An important application of Chebyshev's Inequality is proving the **Weak Law of Large Numbers**. The Weak Law of Large Numbers provides proof of the notion that if n independent and identically distributed random variables, X_1, X_2, \dots, X_n , from a distribution with finite variance are observed, then the sample mean, \bar{X} , should be very close to μ provided n is large. Mathematically, the Weak Law of Large Numbers states that if n independent and identically distributed random variables, X_1, X_2, \dots, X_n are observed from a distribution with finite variance, then, for all $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(\left| \frac{X_1 + \dots + X_n}{n} - \mu \right| \geq \epsilon \right) = 0. \quad (3.18)$$

Proof: Consider the random variables X_1, \dots, X_n such that the mean of each one is μ and the variance of each one is σ^2 . Since

$$E \left[\frac{\sum_{i=1}^n X_i}{n} \right] = \mu \quad \text{and} \quad \text{Var} \left[\frac{\sum_{i=1}^n X_i}{n} \right] = \frac{\sigma^2}{n},$$

use version (a) of Chebyshev's Inequality with $k = \epsilon$ to write

$$\mathbb{P} \left(\left| \frac{X_1 + \dots + X_n}{n} - \mu \right| \geq \epsilon \right) \leq \frac{\sigma^2}{n\epsilon^2},$$

which proves (3.18) since

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(\left| \frac{X_1 + \dots + X_n}{n} - \mu \right| \geq \epsilon \right) \leq \lim_{n \rightarrow \infty} \frac{\sigma^2}{n\epsilon^2} = 0.$$

3.4.8 Skewness

Earlier it was discussed that the second moment about the mean of a random variable X is the same thing as the variance of X . Now, the third moment about the mean of a random variable X is used in the definition of the skewness of X . To facilitate the notation used with skewness, first define a **standardized** random variable X^* to be:

$$X^* = \frac{X - \mu}{\sigma},$$

where μ is the mean of X and σ is the standard deviation of X . Using the standardized form of X , it is easily shown that $E[X^*] = 0$ and $\text{Var}[X^*] = 1$. Define the skewness of a random variable X , denoted γ_1 , to be the third moment about the origin of X^* :

$$\gamma_1 = E[(X^*)^3] = \frac{E[(X - \mu)^3]}{\sigma^3} \quad (3.19)$$

Positive values for γ_1 indicate a distribution that is skewed to the right while negative values for γ_1 indicate a distribution that is skewed to the left. If the distribution of X is symmetric with respect to its mean, then its skewness is zero. That is, $\gamma_1 = 0$ for distributions that are symmetric about their mean. Examples of distributions with various γ_1 coefficients are shown in Figure 3.8.

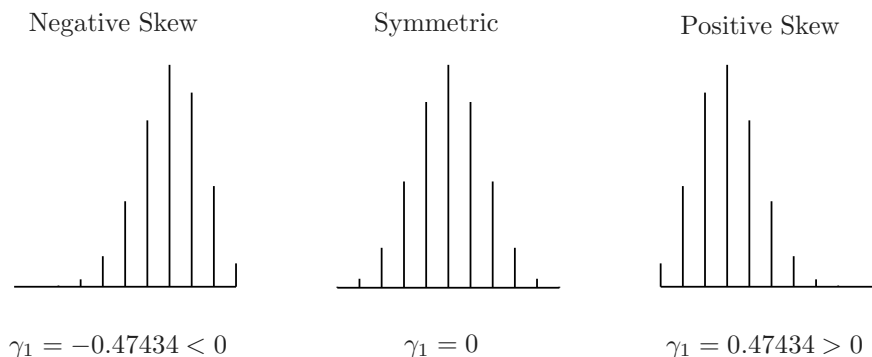


FIGURE 3.8: Distributions with γ_1 (skewness) coefficients that are negative, zero, and positive, respectively.

Example 3.25 Let the pdf of X be defined by $p(x) = x/15$, $x = 1, 2, 3, 4, 5$. Compute γ_1 for the given distribution.

Solution: The value of γ_1 is computed to be

$$\gamma_1 = E[(X^*)^3] = \frac{E[(X - \mu)^3]}{\sigma^3} = -0.5879747$$

which means the distribution has a negative skew. To compute the answer with S, the following facts are used:

1. $\mu = E[X]$.
2. $\sigma = \sqrt{E[X^2] - E[X]^2}$.
3. $X^* = \frac{X - \mu}{\sigma}$.
4. $\gamma_1 = E[(X^*)^3]$.

```
> x <- 1:5
> px <- x/15
> EX <- sum(x*px)
> sigmaX <- sqrt(sum(x^2*px) - EX^2)
```

```
> X.star <- (x-EX)^3/sigmaX^3
> skew <- sum(X.star*px)
> skew
[1] -0.5879747
```



3.4.9 Moment Generating Functions

Finding the first, second, and higher moments about the origin using the definition $\alpha_r = E[X^r]$ is not always an easy task. However, one may define a function of a real variable t called the moment generating function, **mgf**, that can be used to find moments with relative ease provided the **mgf** exists. Given a random variable X with **pdf** $p(x)$, the **mgf** of X , written $M_X(t)$, is defined as

$$M_X(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f(x) dx, \quad -h < t < h. \quad (3.20)$$

provided there is a positive number h such that, for $-h < t < h$, the expectation of e^{tX} exists. If X is discrete, then $E[e^{tX}] = \sum_x e^{tx} p(x)$. When the **mgf** exists, it is unique and completely determines the distribution of the random variable. Consequently, if two random variables have the same **mgf**, they have the same distribution.

Example 3.26 Given the function

$$f(x) = k, \quad -1 < x < 1$$

of the random variable X , find the **mgf** of the distribution using (3.20).

Solution: The reader may verify that a value of $k = \frac{1}{2}$ produces a valid **pdf**. The **mgf** of the distribution will then be

$$\begin{aligned} M_X(t) &= E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f(x) dx, \quad -h < t < h \\ &= \int_{-1}^1 e^{tx} \frac{1}{2} dx = \frac{e^{tx}}{2t} \Big|_{-1}^1 \\ &= \frac{e^t - e^{-t}}{2t}, \quad t \neq 0. \end{aligned}$$

Note that if $t = 0$, then $M_X(t) = 1$ since $M_X(t) = E[e^{tX}] = E[e^0] = 1$. Therefore, the **mgf** is written

$$M_X(t) = \begin{cases} \frac{e^t - e^{-t}}{2t} & \text{if } t \neq 0 \\ 1 & \text{if } t = 0. \end{cases}$$

Theorem 3.2 If X has **mgf** $M_X(t)$, then the derivatives of $M_X(t)$ of all orders exist at $t = 0$, and

$$E[X^r] = \frac{d^r}{dt^r} M_X(t) \Big|_{t=0}.$$

A proof of the last theorem is beyond the scope of this text. However, assuming the distribution is discrete and summation and differentiation may be interchanged, note that

$$\begin{aligned} E[X^1] &= \frac{d^1}{dt^1} M_X(t)|_{t=0} = \frac{d^1}{dt^1} \sum_x e^{tx} p(x)|_{t=0} = \sum_x \frac{d^1}{dt^1} e^{tx} p(x)|_{t=0} \\ &= \sum_x x e^{tx} p(x)|_{t=0} = \sum_x x p(x) = \alpha_1 = E[X^1] \\ E[X^2] &= \frac{d^2}{dt^2} M_X(t)|_{t=0} = \frac{d^2}{dt^2} \sum_x e^{tx} p(x)|_{t=0} = \sum_x \frac{d^2}{dt^2} e^{tx} p(x)|_{t=0} \\ &= \sum_x x^2 e^{tx} p(x)|_{t=0} = \sum_x x^2 p(x) = \alpha_2 = E[X^2] \\ E[X^r] &= \frac{d^r}{dt^r} M_X(t)|_{t=0} = \frac{d^r}{dt^r} \sum_x e^{tx} p(x)|_{t=0} = \sum_x \frac{d^r}{dt^r} e^{tx} p(x)|_{t=0} \\ &= \sum_x x^r e^{tx} p(x)|_{t=0} = \sum_x x^r p(x) = \alpha_r = E[X^r] \end{aligned}$$

Example 3.27 Let X be a random variable with probability distribution

$$P(X = x|n, \pi) = \frac{n!}{(n-x)!x!} \pi^x (1-\pi)^{(n-x)} \quad x = 0, 1, \dots, n$$

Using the moment generating function, check that $E[X] = n\pi$ and $\text{Var}[X] = n\pi(1-\pi)$. (Hint: $(a+b)^n = \sum_{x=0}^n \binom{n}{x} b^x a^{n-x}$.)

Solution: First, the moment generating function is calculated:

$$\begin{aligned} M(t) &= E[e^{tx}] = \sum_{x=0}^n \binom{n}{x} e^{tx} \pi^x (1-\pi)^{n-x} \\ &= \sum_{x=0}^n \binom{n}{x} (\pi e^t)^x (1-\pi)^{n-x} \\ &= [(1-\pi) + \pi e^t]^n \end{aligned}$$

The first and second derivatives of $M(t)$ at $t = 0$ give $E[X]$ and $E[X^2]$, respectively, which are used to calculate the mean and variance of X :

$$M'(t) = n[(1-\pi) + \pi e^t]^{n-1} (\pi e^t)$$

and, using the product and chain rules

$$M''(t) = n(n-1)[(1-\pi) + \pi e^t]^{n-2} (\pi e^t)^2 + n[(1-\pi) + \pi e^t]^{n-1} (\pi e^t).$$

This yields

$$\begin{aligned} E[X] &= M'(0) = n\pi \quad \text{and} \\ \text{Var}[X] &= E[X^2] - E[X]^2 = M''(0) - [M'(0)]^2 = n(n-1)\pi^2 + n\pi - (n\pi)^2 = n\pi(1-\pi) \end{aligned}$$



Theorem 3.3 If a and b are real-valued constants, then

$$(1) M_{X+a}(t) = E [e^{(X+a)t}] = e^{at} \cdot M_X(t).$$

$$(2) M_{bX}(t) = E (e^{bXt}) = M_X(bt).$$

$$(3) M_{\frac{X+a}{b}}(t) = E \left[e^{\left(\frac{X+a}{b}\right)t} \right] = e^{\frac{a}{b}t} \cdot M_X \left(\frac{t}{b} \right).$$

The proof of Theorem 3.3 is left as an exercise for the reader.

3.5 Problems

1. Three dice are thrown. What fraction of the time does a sum of 9 appear on the faces? What percent of the time does a sum of 10 appear?
2. How many different six-place license plates are possible if the first two places are letters and the remaining places are numbers?
3. How many different six-place license plates are possible (first two places letters, remaining places numbers) if repetition among letters and numbers is not permissible?
4. Susie has 25 books she would like to arrange on her desk. Of the 25 books, 7 are statistics books, 6 are biology books, 5 are English books, 4 are history books, and 3 are psychology books. If Susie arranges her books by subject, how many ways can she arrange her books?
5. A hat contains 20 consecutive numbers (1 to 20). If four numbers are drawn at random, how many ways are there for the largest number to be a 16 and the smallest number to be a 5?
6. A university committee of size 10, consisting of 2 faculty from the college of fine and applied arts, 2 faculty from the college of business, 3 faculty from the college of arts and sciences, and 3 administrators, is to be selected from 6 fine and applied arts faculty, 7 college of business faculty, 10 college of arts and sciences faculty, and 5 administrators. How many committees are possible?
7. How many different letter arrangements can be made from the letters BIOLOGY, PROBABILITY, and STATISTICS, respectively.
8. A doll house must be painted and assembled before it can be given as a gift. If there are 12 equal-sized rooms in the doll house and there is enough white paint for 4 rooms, enough pink paint for 3 rooms, and enough blue paint for 5 rooms, in how many ways can the 12 rooms be painted?
9. A shipment of 50 laptops includes 3 that are defective. If an instructor purchases 4 laptops from the shipment to use in his class, how many ways are there for the instructor to purchase at least 2 of the defective laptops?
10. A multiple-choice test consists of 10 questions. Each question has 5 answers (only one is correct). How many different ways can a student fill out the test?
11. How many ways can five politicians stand in line? In how many ways can they stand in line if two of the politicians refuse to stand next to each other?
12. There are five different colored jerseys worn throughout the Tour de France. The yellow jersey is worn by the rider with the least accumulated time; the green jersey is worn by the best sprinter; the red and white polka dot jersey is worn by the best climber. The white jersey is worn by the best youngest rider, and the red jersey is worn by the rider with the most accumulated time still in the race. If 150 riders finish the Tour, how many different ways can the yellow, green, and red and white polka dot jerseys be awarded if (a) a rider can receive any number of jerseys and (b) each rider can receive at most one jersey.

13. A president, treasurer, and secretary, all different, are to be chosen from among the 10 active members of a university club. How many different choices are possible if
 - (a) There are no restrictions.
 - (b) A will serve only if she is the treasurer.
 - (c) B and C will not serve together.
 - (d) D and E will serve together or not at all.
 - (e) F must be an officer.
14. On a multiple-choice exam with three possible answers for each of the five questions, what is the probability that a student would get four or more correct answers just by guessing?
15. Suppose four balls are chosen at random without replacement from an urn containing six black balls and four red balls. What is the probability of selecting two balls of each color?
16. What is the probability that a hand of five cards chosen randomly and without replacement from a standard deck of 52 cards contains the ace of hearts, exactly one other ace, and exactly two kings?
17. Verify that $\mathbb{P}(F|E)$ satisfies the three axioms of probability on page 81.
18. Prove Theorem 3.3 on page 106.
19. In the New York State lottery game, six of the numbers 1 through 54 are chosen by a customer. Then, in a televised drawing, six of these numbers are selected. If all six of a customer's numbers are selected, then that customer wins a share of the first prize. If five or four of the numbers are selected, the customer wins a share of the second or the third prize. What is the probability that any customer will win a share of the first prize, the second prize, and the third prize, respectively?
20. Assume that $\mathbb{P}(A) = 0.5$, $\mathbb{P}(A \cap C) = 0.2$, $\mathbb{P}(C) = 0.4$, $\mathbb{P}(B) = 0.4$, $\mathbb{P}(A \cap B \cap C) = 0.1$, $\mathbb{P}(B \cap C) = 0.2$, and $\mathbb{P}(A \cap B) = 0.2$. Calculate the following probabilities:
 - (a) $\mathbb{P}(A \cup B \cup C)$
 - (b) $\mathbb{P}(A^c \cap (B \cup C))$
 - (c) $\mathbb{P}((B \cap C)^c \cup (A \cap B)^c)$
 - (d) $\mathbb{P}(A) - \mathbb{P}(A \cap C)$
21. Let the random variable X be the sum of the numbers on two fair dice. Find an upper bound on $\mathbb{P}(|X - 7| \geq 4)$ using Chebyshev's Inequality as well as the exact probability for $\mathbb{P}(|X - 7| \geq 4)$.
22. A new drug test being considered by the International Olympic Committee can detect the presence of a banned substance when it has been taken by the subject in the last 90 days 98% of the time. However, the test also registers a "false positive" in 2% of the population that has never taken the banned substance. If 2% of the athletes in question are taking the banned substance, what is the probability a person that has a positive drug test is actually taking the banned substance?

23. The products of an agricultural firm are delivered by four different transportation companies, A, B, C, and D. Company A transports 40% of the products; company B, 30%; company C, 20%; and, finally, company D, 10%. During transportation, 5%, 4%, 2%, and 1% of the products spoil with companies A, B, C, and D, respectively. If one product is randomly selected,
- Obtain the probability that it is spoiled.
 - If the chosen product is spoiled, derive the probability that it has been transported by company A.
24. Two lots of large glass beads are available (A and B). Lot A has four beads, two of which are chipped; and lot B has five beads, two of which are chipped. Two beads are chosen at random from lot A and passed to lot B. Then, one bead is randomly selected from lot B. Find:
- The probability that the selected bead is chipped.
 - The probability that the two beads selected from lot A were not chipped if the bead selected from lot B is not chipped.
25. A box contains 5 defective bulbs, 10 partially defective (they start to fail after 10 hours of use), and 25 perfect bulbs. If a bulb is tested and it does not fail immediately, find the probability that the bulb is perfect.
26. A salesman in a department store receives household appliances from three suppliers: I, II, and III. From previous experience, the salesman knows that 2%, 1%, and 3% of the appliances from supplier I, II, and III, respectively, are defective. The salesman sells 35% of the appliances from supplier I, 25% from supplier II, and 40% from supplier III. If an appliance randomly selected is defective, find the probability that it comes from supplier III.
27. A garage has two machines, A and B, to balance the wheels of a car. Suppose that 95% of the wheels are correctly balanced by machine A, while 85% of the wheels are correctly balanced by machine B. A machine is randomly selected to balance 20 wheels, and 3 of them are not properly balanced. What is the probability that machine A was used? What is the probability machine B was used?
28. An urn contains 14 balls; 6 of them are white, and the others are black. Another urn contains 9 balls; 3 are white, and 6 are black. A ball is drawn at random from the first urn and is placed in the second urn. Then, a ball is drawn at random from the second urn. If this ball is white, find the probability that the ball drawn from the first urn was black.
29. An office supply store is selling packages of 100 CDs at a very affordable price. However, roughly 10% of all packages are defective. If a package of 100 CDs containing exactly 10 defective CDs is purchased, find the probability that exactly 2 of the first 5 CDs used are defective.
30. A box contains six marbles, two of which are black. Three are drawn with replacement. What is the probability two of the three are black?
31. The ASU triathlon club consists of 11 women and 7 men. What is the probability of selecting a committee of size four with exactly three women?

32. Four golf balls are to be placed in six different containers. One ball is red; one, green; one, blue; and one, yellow.
- In how many ways can the four golf balls be placed into six different containers? Assume that any container can contain any number of golf balls (as long as there are a total of four golf balls).
 - In how many ways can the golf balls be placed if container one remains empty?
 - In how many ways can the golf balls be placed if no two golf balls can go into the same container?
 - What is the probability that no two golf balls are in the same container, assuming that the balls are randomly tossed into the containers?
33. Previous to the launching of a new flavor of yogurt, a company has conducted taste tests with four new flavors: lemon, strawberry, peach, and cherry. It obtained the following probabilities of a successful launch: $\mathbb{P}(\text{lemon}) = 2/10$, $\mathbb{P}(\text{strawberry}) = 3/10$, $\mathbb{P}(\text{peach}) = 4/10$, and $\mathbb{P}(\text{cherry}) = 5/10$. Let X be the random variable “number of successful flavors launched.” Obtain its probability mass function.
34. A family has three cars, all with electric windows. Car A’s windows always work. Car B’s windows work 30% of the time, and Car C’s windows work 75% of the time. The family uses Car A $2/3$ of the time; Car B, $2/9$ of the time; and Car C, the remaining fraction.
- On a particularly hot day, when the family wants to roll the windows down, compute the probability the windows will work.
 - If the electric windows work, find the probability the family is driving Car C.
35. John and Peter play a game with a coin such that $\mathbb{P}(\text{head}) = p$. The game consists of tossing a coin twice. John wins if the same result is obtained in the two tosses, and Peter wins if the two results are different.
- At what value of p is neither of them favored by the game?
 - If p is different from your answer in (a), who is favored?
36. A bank is going to place a security camera in the ceiling of a circular hall of radius r . What is the probability that the camera is placed nearer the center than the outside circumference if the camera is placed at random?
37. Anthony and Mark make a bet at the beginning of the school year. If Anthony passes one exam, Mark will pay him €10, but if Anthony fails the exam, he will give €10 to Mark. If Anthony takes 10 exams and the probability of passing an exam is 0.5, find the probability that
- Anthony wins €60.
 - Anthony wins €30.
38. Louis and Joseph have decided to play a beach volleyball match. Each of them put €50 into a pot, so the winner will get €100. The first one to reach 21 points wins. When the score was 19 points for Louis and 18 for Joseph, the match was rained out, and they decided to share the prize so that each one received winnings proportional to the probability of winning the match given their current points. How much money did each receive?

39. Consider tossing three well-made coins. The eight possible outcomes are

$$HHH, HHT, HTH, HTT, THH, THT, TTH, TTT.$$

Define X as the random variable “number of heads showing when three coins are tossed.” Obtain the mean and the variance of X . Simulate tossing three fair coins 10,000 times. Compute the simulated mean and variance of X . Are the simulated values within 2% of the theoretical answers?

40. Every month, a family must decide what to do on Sundays. If they stay at home, they do one of two things with equal probability: have lunch in a restaurant, which costs €100, or go to the park, which is free. Assuming four weeks in a month, compute the probability distribution of expenditures.
41. In a lottery game, one can win €10,000 with probability 0.01 and €1000 with probability 0.05. How much should one pay for a lottery ticket to make the game fair?
42. To play a game, one must bet €100 every time, and the probability of winning €100 is $1/2$. Every day, a person plays uninterruptedly until he loses once. Then, he leaves the game.

- (a) Find the probability that he plays more than four times in one day.
 (b) Find the probability that one day he leaves the game having won €600.
 (c) Calculate the expected winnings per day.

43. Consider the random variable X , which takes the values 1, 2, 3, and 4 with probabilities 0.2, 0.3, 0.1, and 0.4, respectively. Calculate $E[X]$, $1/E[X]$, $E[1/X]$, $E[X^2]$, and $E[X]^2$, and check empirically that $E[X]^2 \neq E[X^2]$ and $E[1/X] \neq 1/E[X]$.
44. Show that the following distribution is a probability mass function. Construct a plot of the probability mass function and obtain the cumulative probability function.

$$\begin{aligned} \mathbb{P}(X = -2) &= 0.2, & \mathbb{P}(1 < X \leq 3) &= 0.1, & \mathbb{P}(X = 4) &= 0.2, \\ \mathbb{P}(5 < X \leq 5.5) &= 0.2, & \mathbb{P}(X = 6) &= 0.15, & \mathbb{P}(7 < X \leq 8) &= 0.15 \end{aligned}$$

45. Two stockbrokers on the floor of the New York Stock Exchange, Alvin and Bob, are interested in purchasing shares from a single company. In a given day, Alvin or Bob buys shares with probability p . Assume that Alvin starts the buying process; when he finishes, Bob is allowed to buy, and so on.

- (a) Find the probability that Alvin buys shares on a given day.
 (b) If two lots of shares are purchased, find the probability that they have been purchased by the same stockbroker.

(Hint: $\sum_{i=0}^{\infty} r^i = \frac{1}{1-r}$ is $|r| < 1$.)

46. Given the function

$$f(x) = k, \quad -1 < x < 1$$

of the random variable X , find the coefficient of skewness for the distribution using (3.19).

47. Consider an experiment where two dice are rolled. Let the random variable X equal the sum of the two dice and the random variable Y be the difference of the two dice.
- Find the mean of X .
 - Find the variance of X .
 - Find the skewness of X .
 - Find the mean of Y .
 - Find the variance of Y .
 - Find the skewness of Y .
48. The number of hits on a faculty member's homework solutions page has an average of 100 hits per day.
- Give an upper bound for the probability the faculty member's homework solutions page has more than 112 hits per day.
 - Suppose the variance of the number of hits is known to be 36. Now, give an upper bound for the probability the faculty member's homework solutions page has more than 112 hits per day.
 - The probability that the number of hits is between 88 and 112 inclusive must be at least what?
 - How many days must visits to the site be recorded so that the average number of hits is within 6 of 100 with a probability of at least 0.9?
49. Find the values of k such that the following functions are probability density functions:
- $f(x) = kx^4/5, 0 < x < 1$.
 - $f(x) = kx^2, 0 < x < 2$.
 - $f(x) = k\sqrt{x}/2, 0 < x < 1$.

Construct plots of these functions and their corresponding cumulative density functions.

50. Given the following cumulative density function,

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x^2}{4} & 0 \leq x \leq 2 \\ 1 & 2 < x, \end{cases}$$

derive the probability density function $f(x)$. Calculate the median of the distribution.

51. Consider the following function:

$$f(x) = \frac{2}{25}(x - 5), \quad 5 \leq x \leq 10.$$

- Show that $f(x)$ satisfies properties 1 and 2 on page 93 of a continuous probability density function.
- Plot $f(x)$.
- Derive and plot $f(x)$'s cumulative probability function, $F(x)$.

- (d) Calculate $\mathbb{P}(X \leq 8)$, $\mathbb{P}(X \geq 6)$, and $\mathbb{P}(7 \leq X \leq 8)$ by hand.
- (e) Calculate $\mathbb{P}(X \leq 8)$, $\mathbb{P}(X \geq 6)$, and $\mathbb{P}(7 \leq X \leq 8)$ using the function `integrate()`.
52. The number of bottles of milk that a dairy farm fills per day is a random variable with mean 5000 and standard deviation 100. Assume the farm always has a sufficient number of glass bottles to be used to store the milk. However, for a bottle of milk to be sent to a grocery store, it must be hermetically sealed with a metal cap that is produced on site. Calculate the minimum number of metal caps that must be produced on a daily basis so that all filled milk bottles can be shipped to grocery stores with a probability of at least 0.9.
53. Define X as the space occupied by certain device in a 1 m^3 container. The probability density function of X is given by

$$f(x) = \frac{630}{56}x^4(1-x^4), \quad 0 < x < 1.$$

- (a) Graph the probability density function.
- (b) Calculate the mean of X by hand.
- (c) Calculate the variance X by hand.
- (d) Calculate $\mathbb{P}(0.20 < X < 0.80)$ by hand.
- (e) Calculate the mean of X using `integrate()`.
- (f) Calculate the variance of X using `integrate()`.
- (g) Calculate $\mathbb{P}(0.20 < X < 0.80)$ using `integrate()`.
54. Consider the probability density function

$$f(x) = \frac{1}{36}xe^{-x/6}, \quad x > 0.$$

Derive the moment generating function, and calculate the mean and the variance.

Chapter 4

Univariate Probability Distributions

4.1 Introduction

This chapter examines univariate (single variable) probability distributions that are used frequently to model random phenomena. Discrete probability distributions are introduced first, followed by continuous probability distributions. Discrete distributions can be used to model the number of failures until a successful rocket launch, the number of passing students in a class, or the number of taxis that pass a street corner, as well as many other phenomena with countable outcomes. Continuous distributions are used to model measurement variables such as weight, height, and time. Joint distributions will be introduced in Chapter 5.

4.2 Discrete Univariate Distributions

4.2.1 Discrete Uniform Distribution

The random variable X is said to follow a discrete uniform distribution with parameter n (where $n \in \mathbb{N}$) if the probability X takes on the value x is the same for all x , where $x = x_1, x_2, \dots, x_n$:

Discrete Uniform Distribution

$$\mathbb{P}(X = x_i | n) = \frac{1}{n}, \quad i = 1, 2, \dots, n.$$

$$E[X] = \frac{1}{n} \sum_{i=1}^n x_i \tag{4.1}$$

$$\text{Var}[X] = \frac{1}{n} \sum_{i=1}^n (x_i - E[X])^2$$

$$M_X(t) = \frac{1}{n} \sum_{i=1}^n e^{tx_i}$$

When $x_i = i$ for $i = 1, \dots, n$, it can be shown that $E[X] = \frac{n+1}{2}$ and that $\text{Var}[X] = \frac{n^2-1}{12}$, respectively.

Example 4.1 One light bulb is randomly selected from a box that contains a 40 watt light bulb, a 60 watt light bulb, a 75 watt light bulb, a 100 watt light bulb, and a 120 watt light bulb. Write the probability function for the random variable that represents the wattage of the randomly selected light bulb, and determine the mean and variance of that random variable.

Solution: The random variable X can assume the set of values $\Omega = \{40, 60, 75, 100, 120\}$. The probability density function for the random variable X is

$$\mathbb{P}(X = x|5) = 1/5 \quad \text{for} \quad x = 40, 60, 75, 100, 120.$$

The expected value of X , $E[X] = 79$, and the variance of X , $Var[X] = 804$. S can be used to alleviate the arithmetic:

```
> Watts <- c(40,60,75,100,120)
> meanWatts <- (1/5)*sum(Watts)
> varWatts<- (1/5)*sum((Watts-meanWatts)^2)
> ans <- c(meanWatts, varWatts)
> ans
[1] 79 804
```

■

4.2.2 Bernoulli and Binomial Distributions

When the same coin is tossed n times by the same person under the same experimental conditions, it stands to reason that each toss of the coin will result in one of two outcomes (heads or tails), that the outcome on any given trial will not influence the outcome of any other trial, and that the probability of getting a head assuming a fair coin on any trial is a constant $\frac{1}{2}$. Tossing a coin a single time is an example of a **Bernoulli** trial. A Bernoulli trial is a random experiment with only two possible outcomes. The outcomes are mutually exclusive and exhaustive, for example, success or failure, true or false, alive or dead, male or female, etc. A Bernoulli random variable, X , can take on two values, where $X(\text{success}) = 1$ and $X(\text{failure}) = 0$. The probability that X is a success is π , and the probability that X is a failure is $\varrho = 1 - \pi$. The **pdf**, mean, variance, and **mgf** of a Bernoulli random variable are in (4.2).

<p>Bernoulli Distribution $X \sim \text{Bernoulli}(\pi)$</p> $\mathbb{P}(X = x \pi) = \pi^x(1 - \pi)^{1-x}, x = 0, 1$ $E[X] = \pi$ $Var[X] = \pi(1 - \pi)$ $M_X(t) = \pi e^t + \varrho$	(4.2)
--	-------

When a sequence of Bernoulli trials conforms to the following list of requirements it is called a **binomial experiment**:

1. The experiment consists of a fixed number (n) of Bernoulli trials.
2. The probability of success for each trial, denoted by π , is constant from trial to trial. The probability of failure is $\varrho = (1 - \pi)$.

3. The trials are independent.
4. The random variable of interest, X , is the number of observed successes during the n trials.

The probability that X is equal to x can be found in the following fashion. Any particular sequence of x successes occurs with probability $\pi^x(1-\pi)^{(n-x)}$ since there are x successes and $(n-x)$ failures. However, there are $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ possible sequences of x successes. Write $X \sim \text{Bin}(n, \pi)$ to indicate the random variable X follows a binomial distribution with parameters n and π . The probability X is equal to x , the mean, the variance, and the moment generating function of a binomial random variable are in (4.3).

<p>Binomial Distribution $X \sim \text{Bin}(n, \pi)$</p> $\mathbb{P}(X = x n, \pi) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}, \quad x = 0, 1, 2, \dots, n.$ $E[X] = n\pi$ $\text{Var}[X] = n\pi(1 - \pi)$ $M_X(t) = (\pi e^t + \varrho)^n$	(4.3)
--	-------

It is left as an exercise for the student to verify that $E[X] = n\pi$, $\text{Var}[X] = n\pi(1 - \pi)$, and that the moment generating function of a binomial random variable is $M_X(t) = (\pi e^t + \varrho)^n$. (See Problem 40 on page 167.)

Code to create graphs that represent the probability density function and the cumulative distribution function for a $\text{Bin}(8, 0.3)$ random variable follows. The graphs that are created are similar to those in Figure 4.1 on the next page.

```
> par(mfrow=c(1,2), pty="s")
> plot(0:8, dbinom(0:8,8,0.3), type="h", xlab="x", ylab="P(X=x)",
+ xlim=c(-1,9))
> title("PDF for X~Bin(8, 0.3)")
> plot(0:8, pbinom(0:8,8,0.3), type="n", xlab="x", ylab="F(x)",
+ xlim=c(-1,9), ylim=c(0,1))
> segments(-1,0,0,0)
> segments(0:8, pbinom(0:8,8,.3), 1:9, pbinom(0:8,8,.3))
> lines(0:7, pbinom(0:7,8,.3), type="p", pch=16)
> segments(-1,1,9,1, lty=2)
> title("CDF for X~Bin(8, 0.3)")
```

Example 4.2 ▷ *Simulating Bernoulli* ◁ Write a function that will generate m repeated samples of n Bernoulli trials each with probability of success π . Use the function to generate 1000 samples of size $n = 5$ with $\pi = 0.5$ to simulate the binomial distribution. Have the function create frequency tables for both the simulated and the theoretical random variable so that comparisons can be made between the two. Finally, produce a histogram of the simulated successes with the theoretical probability for the random variable X that has a binomial distribution with $n = 5$ and $\pi = 0.5$ superimposed over the simulated values.

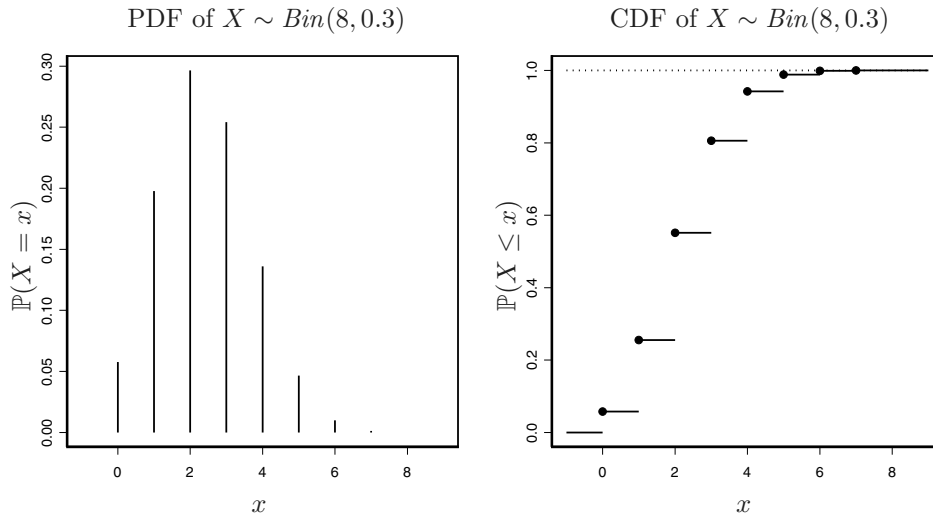


FIGURE 4.1: Left graph is the probability density function (**pdf**) of a binomial random variable with $n = 8$ and $\pi = 0.3$. Right graph is the cumulative distribution function (**cdf**) of a binomial random variable with $n = 8$ and $\pi = 0.3$.

Solution: The function `binogen()` is written to solve Example 4.2 in general:

```
binogen <- function(samples, n, pi) {
  values <- sample(c(0,1), samples*n, replace=TRUE, prob=c(pi,1-pi))
  value.mat <- matrix(values, ncol=n)
  Successes <- apply(value.mat, 1, sum)
  a1 <- round((table(Successes)/samples), 3)
  b1 <- round(dbinom(0:n, n, 1-pi), 3)
  names(b1) <- 0:n
  hist(Successes, breaks=c((-0.5+0):(n+0.5)), probability=TRUE,
       ylab="", main=" Theoretical Values Superimposed
       Over Histogram of Simulated Values", col=13)
  x <- 0:n
  fx <- dbinom(x, n, 1-pi)
  lines(x, fx, type="h")
  lines(x, fx, type="p", pch=16)
  list(simulated.distribution=a1, theoretical.distribution=b1)}
```

Then, the results from using the function to generate 1000 samples where $n = 5$ and $\pi = 0.5$ answer Example 4.2 in particular:

```
> binogen(1000, 5, 0.5)
$simulated.distribution
Successes
  0    1    2    3    4    5
0.023 0.174 0.311 0.308 0.153 0.031

$theoretical.distribution
  0    1    2    3    4    5
0.031 0.156 0.312 0.312 0.156 0.031
```

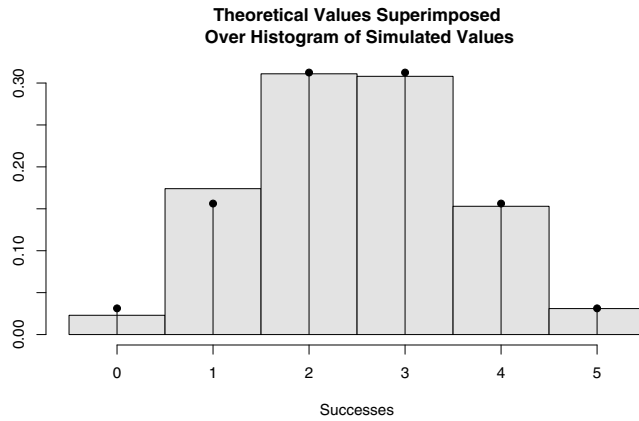


FIGURE 4.2: Histogram of 1000 simulated samples where $n = 5$ and $\pi = 0.5$ superimposed on the theoretical distribution for a random variable following a $Bin(5, 0.5)$ distribution.

Using the function `rbinom()`, one can generate 1000 samples of a $Bin(n = 5, \pi = 0.5)$ distribution by entering

```
> x <- rbinom(1000, 5, .5)
> table(x)/1000      # Empirical distribution
x
  0     1     2     3     4     5
0.042 0.145 0.302 0.313 0.163 0.035
```

If one wants to generate the same numbers at a later date, the command `set.seed()` can be used. The graph in Figure 4.2 was created with `set.seed(31)`. ■

Example 4.3 ▷ *Binomial Calculation* ◁ Consider the problem of calculating the probability of obtaining 6 or more heads in 10 tosses of a weighted coin, where the probability of obtaining a head in any given trial is 0.33.

Solution: Let the random variable X equal the number of trials that result in a head. Consequently, $X \sim Bin(10, 0.33)$, and the sum of the individual probabilities of obtaining 6, 7, 8, 9, and 10 heads needs to be found. Mathematically, this is written $\mathbb{P}(X \geq 6) = \mathbb{P}(X = 6) + \mathbb{P}(X = 7) + \cdots + \mathbb{P}(X = 10)$, where

$$\begin{aligned} \mathbb{P}(X = 6) &= \frac{10!}{6!(10-6)!} \times 0.33^6 \times (1 - 0.33)^{(10-6)} = 0.0546515 \\ \mathbb{P}(X = 7) &= \frac{10!}{7!(10-7)!} \times 0.33^7 \times (1 - 0.33)^{(10-7)} = 0.0153817 \\ \mathbb{P}(X = 8) &= \frac{10!}{8!(10-8)!} \times 0.33^8 \times (1 - 0.33)^{(10-8)} = 0.0028410 \\ \mathbb{P}(X = 9) &= \frac{10!}{9!(10-9)!} \times 0.33^9 \times (1 - 0.33)^{(10-9)} = 0.0003110 \\ \mathbb{P}(X = 10) &= \frac{10!}{10!(10-10)!} \times 0.33^{10} \times (1 - 0.33)^{(10-10)} = 0.0000153 \end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{P}(X \geq 6) &= \mathbb{P}(X = 6) + \mathbb{P}(X = 7) + \mathbb{P}(X = 8) + \mathbb{P}(X = 9) + \mathbb{P}(X = 10) \\ &= 0.0546515 + 0.0153817 + 0.0028410 + 0.0003110 + 0.0000153 \\ &= 0.0732005\end{aligned}$$

There are several approaches one might take to solve the problem with `S`. One should realize that the following are all equivalent statements:

$$\begin{aligned}\mathbb{P}(X \geq 6) &= \mathbb{P}(X = 6) + \mathbb{P}(X = 7) + \mathbb{P}(X = 8) + \mathbb{P}(X = 9) + \mathbb{P}(X = 10) \\ &= 1 - \mathbb{P}(X \leq 5) \\ &= 1 - [\mathbb{P}(X = 5) + \mathbb{P}(X = 4) + \cdots + \mathbb{P}(X = 0)].\end{aligned}$$

To find $\mathbb{P}(X \geq 6)$ with `S`, compute the individual probabilities with `dbinom(6:10,10,.33)` and then sum them with the command `sum()` by typing `sum(dbinom(6:10,10,.33))`. Another solution is to find $1 - \mathbb{P}(X \leq 5)$, which is accomplished with `1-pbinom(5,10,.33)` or `1-sum(dbinom(5:0,10,.33))`. Note that `dbinom()` computes $\mathbb{P}(X = x)$, the **pdf**, while `pbinom()` gives $\mathbb{P}(X \leq x)$, the **cdf**.

```
> sum(dbinom(6:10,10,0.33))
[1] 0.07320046
> 1 - pbinom(5,10,0.33)
[1] 0.07320046
> 1 - sum(dbinom(5:0,10,0.33))
[1] 0.07320046
```

■

4.2.3 Poisson Distribution

The Poisson distribution is very popular for modeling the number of times particular events occur in given times or on defined spaces. For example, one might count the number of phone calls to 911 between 1 A.M. and 2 A.M., the number of accidents at a busy street corner during a 24 hour period, or the number of typographical errors on a single page of this book. Unfortunately, the derivation of the Poisson distribution is not straightforward. Instead of deriving the Poisson distribution directly, it is shown that the limiting distribution of the binomial distribution is the Poisson distribution. Actual derivation of the Poisson distribution function is beyond the scope of the current text.

When the number of outcomes in a given continuous interval are counted, an approximate **Poisson process** with parameter $\lambda > 0$ results if the following conditions are satisfied:

- (1) The number of outcomes in non-overlapping intervals are independent. In other words, the number of outcomes in the interval of time $(0, t]$ are independent from the number of outcomes in the interval of time $(t, t + h]$ for any $h > 0$.
- (2) The probability of two or more outcomes in a sufficiently short interval is virtually zero. In other words, provided h is sufficiently small, the probability of obtaining two or more outcomes in the interval $(t, t + h]$ is negligible compared to the probability of obtaining one or zero outcomes in the same interval of time.
- (3) The probability of exactly one outcome in a sufficiently short interval or small region is proportional to the length of the interval or region. In other words, the probability of one outcome in an interval of length h is λh .

When an experiment satisfies the conditions for the Poisson process, the resulting random variable, X , the number of outcomes, is called a Poisson random variable. The probability distribution of the Poisson random variable X , representing the number of outcomes in a given time interval or space region denoted by t , is

$$\mathbb{P}(X = x|\lambda t) = \frac{e^{-\lambda}(\lambda t)^x}{x!} \quad x = 0, 1, \dots, \quad \lambda > 0. \quad (4.4)$$

Although the Poisson distribution is typically used for problems involving time or space, it can be viewed as the limiting form of the binomial distribution. Suppose there is an experiment that satisfies the three criteria for an approximate Poisson process. Let X represent the number of outcomes in an interval of length 1 ($t = 1$). To find $\mathbb{P}(X = x)$, divide the interval of length 1 into n subintervals of equal length. Provided n is much larger than x , the probability of one outcome in any given interval of length $1/n$ is approximately λ/n by criterion (3) of the Poisson process on the preceding page. Substituting $\pi = \lambda/n$ into the binomial probability distribution gives

$$\begin{aligned} \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} &= \frac{n(n-1)\cdots(n-x+1)\lambda^x}{x! n^x} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x} \\ &= \frac{\lambda^x}{x!} \left[\frac{n}{n} \frac{(n-1)}{n} \cdots \frac{(n-x+1)}{n} \right] \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x}. \end{aligned}$$

Now, if x is fixed and $n \rightarrow +\infty$ and $\pi \rightarrow 0$, so that $\lambda = n\pi$ remains constant, the expression between the braces goes to 1 and $\left(1 - \frac{\lambda}{n}\right)^{-x}$ is also 1. Using the fact that $\lim_{n \rightarrow \infty} (1 - \lambda/n)^n = e^{-\lambda}$, obtain $\frac{\lambda^x e^{-\lambda}}{x!}$. The Poisson distribution can be used to approximate binomial probabilities with $\lambda = n\pi$ provided $\pi \leq 0.1$ and $n\pi \leq 5$. See Example 4.8 on page 126 for an example of how the Poisson distribution is used to approximate the probabilities of a binomial distribution.

<p>Poisson Distribution $X \sim Pois(\lambda)$</p> $\mathbb{P}(X = x \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \dots \quad (4.5)$ $E[X] = \lambda$ $Var[X] = \lambda$ $M_X(t) = e^{\lambda(e^t - 1)}$

Note that the parameter λ , referred to as the intensity parameter, represents the mean number of outcomes in either a fixed time interval or a fixed spatial region. The Poisson distribution is particularly appropriate for modeling “rare” phenomena or outcomes where the probability of success is small. However, whether or not data can be viewed as Poisson data depends on whether the proportions of 0’s, 1’s, 2’s, and so on, are similar to those predicted by the Poisson pdf given in (4.5). Given n independent Poisson random variables X_1, X_2, \dots, X_n with parameters $\lambda_1, \lambda_2, \dots, \lambda_n$, respectively, $Y = \sum_{i=1}^n X_i \sim Pois(\sum_{i=1}^n \lambda_i = \lambda)$.

Example 4.4 ▷ *Poisson: World Cup Soccer* ◁ The World Cup is played once every four years. National teams from all over the world compete. In 2002 and in 1998, 36 teams were invited; whereas, in 1994 and in 1990, only 24 teams participated. The data frame `Soccer` contains three columns: `CGT`, `Game`, and `Goals`. All of the information contained in `Soccer` is indirectly available from the FIFA World Cup website, located at <http://fifaworldcup.yahoo.com/>. The numbers of goals scored in the regulation 90 minute periods of World Cup soccer matches from 1990 to 2002 are listed in column `Goals`. There were a total of 575 goals scored during regulation time. The game in which the goals were scored is in column `Game`. There were 232 World Cup soccer games played from 1990 to 2002. There were 64 games played in each of 2002 and 1998 and 54 games played in each of 1994 and 1990. The cumulative goal time is provided in column `CGT`. For example, the first goal was scored at the 67th minute of the first game and the second goal was scored at the 42nd minute of the second game. Consequently, the times listed in `CGT` for the first two goals are 67, and $132 = 90 + 42$. For consistency, all goals scored during injury time are recorded in either the 45th or 90th minute, depending on the half when the injury occurred. Analyze the number of goals scored during regulation play (90 minutes) of World Cup soccer matches to verify that the scores follow an approximate Poisson distribution (Chu, 2003).

Solution: To investigate whether criterion (1) of the Poisson process on page 120 is reasonable, examine the one, two, three, four, and five game lagged correlation coefficients:

```
> attach(Soccer)
> L1 <- Goals[1:228]
> L2 <- Goals[2:229]
> L3 <- Goals[3:230]
> L4 <- Goals[4:231]
> L5 <- Goals[5:232]
> LAG <- cbind(L1, L2, L3, L4, L5)
> # or more succinctly
> LAG <- sapply( 1:5, function(x){Goals[x:(x+227)]} )
> round(cor(LAG),3)
      [,1] [,2] [,3] [,4] [,5]
[1,] 1.000 -0.049 0.055 -0.138 -0.008
[2,] -0.049 1.000 -0.046 0.044 -0.138
[3,] 0.055 -0.046 1.000 -0.054 0.045
[4,] -0.138 0.044 -0.054 1.000 -0.057
[5,] -0.008 -0.138 0.045 -0.057 1.000
```

Independence seems reasonable due to the small correlation coefficients (near zero) but should also be computed with time periods smaller than 90 minutes. Criterion (2) of the Poisson process on page 120, appears satisfied since two goals are never registered during the same one minute period. One way to investigate this is to create a table of the interarrival goal times and note that 0 is not in the table. Whether criterion (3) of the Poisson process on page 120 is satisfied is addressed in Problem 45 on page 169 at the end of the chapter. Next, examine the data to see how well they conform to the Poisson distribution. To calculate the observed number of goals scored during regulation time for the 232 World Cup soccer matches, use `table()`:

```
> table(Goals)
 0  1  2  3  4  5  6  7  8
19 49 60 47 32 18 3 3 1
```

Since there are NA values in the `Goals` column, use the `na.rm=TRUE` and `na.method="omit"` options for the S functions `mean()` and `var()`, respectively. To verify that the mean and the variance of `Goals` are approximately equal, type

```
> mean(Goals, na.rm=TRUE)
[1] 2.478448
> var(Goals, na.rm=TRUE)      # na.method="omit" for S-PLUS
[1] 2.458408
```

Because the mean and variance of `Goals` are approximately equal, it is reasonable to proceed in analyzing the frequencies of `Goals` in comparison to those of a Poisson distribution with $\lambda = 2.478448$. Create a table to facilitate comparing the observed values (`OBS`) to the expected values (`EXP`) as well as the empirical proportions (`Empir`) to the theoretical proportions (`TheoP`) for a Poisson distribution with $\lambda = 2.478448$, the mean number of goals per game. The empirical proportions are merely the number of goals in each category divided by the total number of goals.

```
> OBS <- table(Goals)
> Empir <- round(OBS/sum(OBS), 3)
> TheoP <- round(dpois(0:(length(OBS)-1), mean(Goals, na.rm=TRUE)), 3)
> EXP <- round(TheoP*232, 0)
> ANS <- cbind(OBS, EXP, Empir, TheoP)
> ANS
  OBS EXP Empir TheoP
0  19  19 0.082 0.084
1  49  48 0.211 0.208
2  60  60 0.259 0.258
3  47  49 0.203 0.213
4  32  31 0.138 0.132
5  18  15 0.078 0.065
6   3   6 0.013 0.027
7   3   2 0.013 0.010
8   1   1 0.004 0.003
> detach(Soccer)
```

Since the observed values are close to the expected values, the empirical proportions will be close to the theoretical probabilities. This, in conjunction with the fact that the sample mean (2.478448) is roughly equal to the sample variance (2.458408), implies that modeling the number of goals scored during World Cup soccer games with a Poisson distribution is reasonable. ■

Code to represent a probability density function and cumulative distribution function for a $Pois(\lambda = 1)$ random variable similar to the one shown in Figure 4.3 on the next page is

```
> par(mfrow=c(1,2), pty="s")
> plot(0:8, dpois(0:8,1), type="h", xlab="x", ylab="P", xlim=c(-1,9),
+ main="PDF")
> plot(0:8, ppois(0:8,1), type="n", xlab="x", ylab="F", xlim=c(-1,9),
+ ylim=c(0,1), main="CDF")
> segments(-1,0,0,0)
> segments(0:8, ppois(0:8,1), 1:9, ppois(0:8,1))
```

```
> lines(0:7, ppois(0:7,1), type="p", pch=16)
> segments(-1,1,9,1, lty=2)
```

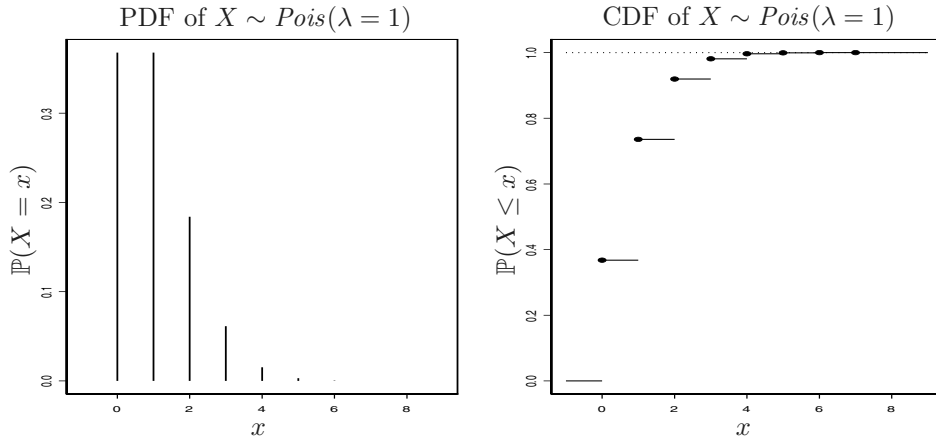


FIGURE 4.3: Left graph is the probability density function (**pdf**) of a Poisson random variable with $\lambda = 1$. Right graph is the cumulative distribution function (**cdf**) of a Poisson random variable with $\lambda = 1$.

Example 4.5 Given a random variable X that follows a Poisson distribution with parameter λ , find the mean and variance of X . Use the fact that

$$e^\lambda = \sum_{r=0}^{\infty} \frac{\lambda^r}{r!} = 1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \cdots .$$

Solution:

$$E[X] = \sum_{r=0}^{\infty} r \frac{\lambda^r}{r!} e^{-\lambda} = \lambda e^{-\lambda} \sum_{r=1}^{\infty} \frac{\lambda^{r-1}}{(r-1)!} = \lambda$$

$$\text{Var}[X] = \sum_{r=0}^{\infty} (r - \lambda)^2 \frac{\lambda^r}{r!} e^{-\lambda}$$

Rearranging terms,

$$\begin{aligned} \text{Var}[X] &= e^{-\lambda} \left\{ \sum_{r=0}^{\infty} r^2 \frac{\lambda^r}{r!} + \sum_{r=0}^{\infty} \lambda^2 \frac{\lambda^r}{r!} - 2\lambda \sum_{r=0}^{\infty} r \frac{\lambda^r}{r!} \right\} \\ &= e^{-\lambda} \left\{ \sum_{r=1}^{\infty} r \frac{\lambda^r}{(r-1)!} + \lambda^2 e^\lambda - 2\lambda \cdot \lambda \sum_{r=1}^{\infty} \frac{\lambda^{r-1}}{(r-1)!} \right\} \\ &= e^{-\lambda} \left\{ \sum_{r=1}^{\infty} (r-1+1) \frac{\lambda^r}{(r-1)!} + \lambda^2 e^\lambda - 2\lambda^2 e^\lambda \right\} \\ &= e^{-\lambda} \left\{ \sum_{r=1}^{\infty} (r-1) \frac{\lambda^r}{(r-1)!} + \sum_{r=1}^{\infty} \frac{\lambda^r}{(r-1)!} + \lambda^2 e^\lambda - 2\lambda^2 e^\lambda \right\} \\ &= e^{-\lambda} \{ \lambda^2 + \lambda + \lambda^2 - 2\lambda^2 \} e^\lambda = \lambda. \end{aligned}$$

■

Example 4.6 More accidents are registered in auto body repair shops during the months of May and June than in the rest of the year. Suppose a particular auto body repair shop has an average of four accidents per month. What is the probability there will be more than seven accidents in this auto body shop during the month of May? What is the probability no more than three accidents will occur during the months of May and June?

Solution: Assuming accidents in the auto body shop follow an approximate Poisson process, the probability of x accidents in one month is

$$\mathbb{P}(X = x) = \frac{4^x e^{-4}}{x!} \quad \text{for } x = 0, 1, 2, \dots$$

The probability more than seven accidents occur during the month of May is

$$\mathbb{P}(X > 7) = 1 - \mathbb{P}(X \leq 7) = 1 - \sum_{i=0}^7 \frac{4^i e^{-4}}{i!} = 0.051.$$

Since the expected number of accidents during May and June is $\lambda' = 2 \cdot 4 = 8$, the probability no more than three accidents occur for the two months in question is calculated as

$$\mathbb{P}(X \leq 3) = \sum_{i=0}^3 \frac{8^i e^{-8}}{i!} = 0.042.$$

The S command to find $1 - \mathbb{P}(X \leq 7)$ is `1-ppois(7,4)`, while $\mathbb{P}(X \leq 3)$ is found by entering `ppois(3,8)`:

```
> 1 - ppois(7,4)
[1] 0.05113362
> ppois(3,8)
[1] 0.04238011
```

■

Example 4.7 Telephone calls to a local 911 number are known to follow a Poisson distribution with an average of two calls per minute. Compute the probability that

- There will be zero calls during a one minute period.
- There will be less than five calls in a one minute period.
- There will be less than six calls in one hour.

Solution: The answers are as follows:

(a) $\mathbb{P}(X = 0; \lambda = 2) = \frac{\lambda^0 e^{-\lambda}}{0!} = \frac{2^0}{0!} e^{-2} = 0.135.$

(b) $\mathbb{P}(X \leq 4; \lambda = 2) = \sum_{r=0}^4 \frac{\lambda^r e^{-\lambda}}{r!} = e^{-2} \left(1 + 2 + \frac{2^2}{2!} + \frac{2^3}{3!} + \frac{2^4}{4!} \right) = 0.947.$

(c) Note that the time period changes from one minute to one hour (60 minutes). Consequently, the average number of calls in one hour is $\lambda' = 2 \times (60) = 120.$

$$\begin{aligned} \mathbb{P}(X \leq 5; \lambda' = 120) &= \sum_{r=0}^5 \frac{\lambda'^r e^{-\lambda'}}{r!} \\ &= e^{-120} \left(1 + 120 + \frac{120^2}{2!} + \frac{120^3}{3!} + \frac{120^4}{4!} + \frac{120^5}{5!} \right) = 0. \end{aligned}$$

The S commands to find $\mathbb{P}(X = 0; \lambda = 2)$, $\mathbb{P}(X \leq 4; \lambda = 2)$, and $\mathbb{P}(X \leq 5; \lambda = 120)$ are `dpois(0,2)`, `ppois(4,2)`, and `ppois(5,120)`, respectively:


```

> dpois(0,2)
[1] 0.1353353
> ppois(4,2)
[1] 0.947347
> ppois(5,120)
[1] 1.658476e-44

```

■

Example 4.8 Numerically show the results from approximating a $Bin(n = 100, \pi = 0.04)$ distribution with a $Pois(\lambda = 4)$.

Solution: The probability distribution function for a $Bin(100, 0.04)$ random variable is

$$\mathbb{P}_{Bin}(X = x) = \binom{100}{x} (0.04)^x (0.96)^{100-x}, \quad x = 0, 1, 2, \dots$$

Since $\pi < 0.1$ and $\lambda = n\pi = 100(0.04) = 4 < 5$, the Poisson distribution can be used to obtain reasonable approximations to the binomial distribution. The probability distribution for a $Pois(4)$ is

$$\mathbb{P}_{Pois}(X = x) = \frac{e^{-4} 4^x}{x!}, \quad x = 0, 1, 2, \dots$$

The first eight values of x for $\mathbb{P}_{Bin}(X = x)$ and $\mathbb{P}_{Pois}(X = x)$ are given in Table 4.1.

Table 4.1: Comparison of binomial and Poisson probabilities

x	0	1	2	3	4	5	6	7	8
$\mathbb{P}_{Bin}(X = x)$	0.017	0.070	0.145	0.197	0.199	0.160	0.105	0.059	0.029
$\mathbb{P}_{Pois}(X = x)$	0.018	0.073	0.147	0.195	0.195	0.156	0.104	0.060	0.030

Note that the results between $\mathbb{P}_{Bin}(X = x)$ and $\mathbb{P}_{Pois}(X = x)$ are virtually identical out to two decimal places. The values in Table 4.1 were generated using S commands as follows:

```

> r <- seq(0,8,1)
> round(dbinom(r,100,0.04), 3)
[1] 0.017 0.070 0.145 0.197 0.199 0.160 0.105 0.059 0.029
> round(dpois(r,4), 3)
[1] 0.018 0.073 0.147 0.195 0.195 0.156 0.104 0.060 0.030

```

■

4.2.4 Geometric Distribution

The geometric distribution, like the binomial distribution, is based on Bernoulli trials. However, the geometric distribution does not fix the number of trials prior to the experiment. The geometric distribution computes the probability the first success occurs after r failures instead of computing the probability of observing x successes in n trials. A random variable X that counts the number of Bernoulli trials that result in failure before the first success is called a **geometric** random variable. Clearly, the probability of a success after r failures is $\pi \times (1 - \pi)^r$, which leads to the geometric probability distribution function where $\varrho = 1 - \pi$ is the probability of failure as it was for the Bernoulli and binomial distributions. The **pdf**, mean, variance, and **mgf** for a geometric random variable are in (4.6).

Geometric Distribution

$$X \sim Geo(\pi)$$

$$\mathbb{P}(X = x; \pi) = \pi \rho^x, \quad x = 0, 1, \dots$$

$$E[X] = \frac{\rho}{\pi} \tag{4.6}$$

$$Var[X] = \frac{\rho}{\pi^2}$$

$$M_X(t) = \frac{\pi}{1 - \rho e^t}$$

Example 4.9 ▷ *Geometric Distribution: Hiring a CPA* ◁ It is known that 20% of all applicants for an overseas position with an international accounting firm speak a foreign language and have passed the CPA (certified public accountant) exam. If applicants are selected at random and interviewed one at a time for the position,

- Compute the probability that the first applicant who speaks a foreign language and has passed the CPA exam is the fourth applicant interviewed.
- Suppose the first applicant that speaks a foreign language who has passed the CPA exam is offered the position and that the applicant accepts the offer. If the accounting firm spends 200 dollars for each interview, what are the expected value and standard deviation of the firm's cost for filling the position.

Solution: The answers are as follows:

(a) Let the random variable X represent the number of applicants interviewed who neither speak a foreign language nor have passed the CPA exam before the first applicant who both speaks a foreign language and has passed the CPA exam is interviewed. The random variable $X \sim Geo(\pi = 0.2)$ and the $\mathbb{P}(X = 3)$ is computed using the **pdf** from (4.6) as

$$\mathbb{P}(X = 3) = \pi \rho^3 = 0.2(0.8)^3 = 0.1024.$$

When $X \sim Geo(\pi = 0.2)$, the $\mathbb{P}(X = 3)$ can be found with **S** using the command `dgeom(3,0.2)`:

```
> dgeom(3,0.2)
[1] 0.1024
```

(b) Be careful with this problem! The expected value and standard deviation of the cost for filling the position are not the same as the expected value and standard deviation of the random variable X as defined in the solution for part (a). Since the question asks for the expected value and standard deviation of the cost for filling the position (r failures and one

success),

$$\begin{aligned}
 E[200(X + 1)] &= 200E[(X + 1)] \\
 &= 200(E[X] + 1) \\
 &= 200\left(\frac{0.8}{0.2} + 1\right) = 1000 \text{ dollars.} \\
 \text{Var}[200(X + 1)] &= 40,000 \text{Var}[(X + 1)] \\
 &= 40,000 \text{Var}[X] \\
 &= 40,000\left(\frac{0.8}{0.2^2}\right) = 800,000 \text{ dollars}^2 \\
 \Rightarrow \sigma_{200(X+1)} &= \sqrt{\text{Var}[200(X + 1)]} = 894.43 \text{ dollars} \quad \blacksquare
 \end{aligned}$$

4.2.5 Negative Binomial Distribution

The geometric random variable counted the number of failures prior to the first success. Quite often, the number of Bernoulli trials required to achieve some fixed number (r) of successes is the problem of interest. When the random variable X is defined as the number of failures prior to the r^{th} success, X has a **negative binomial** distribution written $X \sim NB(r, \pi)$. To find the $\mathbb{P}(X = x)$, first find the probability of $r - 1$ successes in the first $x + r - 1$ trials, and then multiply by the probability of success on the $(x + r)^{\text{th}}$ trial, $\binom{x+r-1}{r-1}\pi^{r-1}(1-\pi)^x \times \pi$. Combining like terms leads to the probability distribution for the negative binomial given in (4.7). The mean, variance, and **mgf** are also in (4.7):

<p>Negative Binomial Distribution $X \sim NB(r, \pi)$</p> $\mathbb{P}(X = x r, \pi) = \binom{x+r-1}{r-1} \pi^r \varrho^x, \quad x = 0, 1, 2, \dots$ $E[X] = r \frac{\varrho}{\pi}$ $\text{Var}[X] = r \frac{\varrho}{\pi^2}$ $M_X(t) = \pi^r (1 - \varrho e^t)^{-r}$	(4.7)
---	-------

Useful Relationships

1. If n independent random variables X_1, \dots, X_n have a geometric distribution with parameter π , then the sum of the n independent random variables follows a negative binomial distribution with parameters (n, π) .
2. If n independent random variables X_1, \dots, X_n have a negative binomial distribution with parameters r_i and π , then the sum of the n random variables is $NB(\sum_{i=1}^n r_i, \pi)$.
3. When $X \sim NB(r, \pi)$ and $r = 1$, a negative binomial random variable is the same as a geometric random variable with parameter π .

Example 4.10 In a particular lot of white wall tires, 10% are missing their white wall. What is the probability one will have to examine six tires before finding four tires with white walls?

Solution: Let the random variable X represent the number of tires without white walls examined before obtaining four tires with white walls. In other words, $X \sim NB(4, 0.9)$ and it follows that

$$\begin{aligned}\mathbb{P}(X = 2|4, 0.9) &= \binom{2+4-1}{4-1} (0.9)^4 (0.1)^2 \\ &= \frac{5!}{3!(2!)} (0.9)^4 (0.1)^2 = 0.066.\end{aligned}$$

To compute the answer in S use the command `dnbinom(x, r, pi)`:

```
> dnbinom(2, 4, 0.9)
[1] 0.06561
```

■

4.2.6 Hypergeometric Distribution

When working with finite populations, the binomial model often becomes untenable. Specifically, when sampling without replacement, the assumption of constant probability from trial to trial is no longer satisfied. However, deriving the exact distribution for a finite sample of dichotomous objects is not difficult. Given a dichotomous population of objects such that m are good and n are bad, the probability of selecting exactly x good items and $k - x$ bad items from a sample of size k is $\binom{m}{x} \binom{n}{k-x} / \binom{m+n}{k}$. Consequently, the random variable X that represents the number of good items selected from a total of m good items in a sample of size k is a **hypergeometric** random variable.

<p>Hypergeometric Distribution $X \sim \text{Hyper}(m, n, k)$</p> $\mathbb{P}(X = x m, n, k) = \frac{\binom{m}{x} \binom{n}{k-x}}{\binom{N}{k}},$ <p>for $x = \max\{0, k - n\}, \dots, \min\{m, k\}$, where $N = m + n$</p> $E[X] = \frac{m \times k}{N}$ $\text{Var}[X] = \frac{m \times n \times k \times (N - k)}{N^2 \times (N - 1)}$	(4.8)
--	-------

One should note that when $\frac{k}{N}$ is small (≤ 0.10), the distribution of a hypergeometric random variable does not differ greatly from the distribution of a binomial random variable with parameters $n = k$ and $\pi = \frac{m}{N}$.

Example 4.11 A computer manufacturer decides to purchase monitors from a new start-up company claiming strict quality control standards. The manufacturer orders 150 monitors and decides to accept the lot provided a random sample of size 25 reveals no defective monitors. If the lot of 150 monitors contains three defective monitors, what is the probability the lot will be accepted?

Solution: Let the random variable X represent the number of non-defective monitors in the sample. Since $X \sim \text{Hyper}(147, 3, 25)$, the $\mathbb{P}(X = 25 | m = 147, n = 3, k = 25)$ is computed as

$$\mathbb{P}(X = 25 | m = 147, n = 3, k = 25) = \frac{\binom{147}{25} \binom{3}{0}}{\binom{150}{25}} = 0.5764.$$

To compute the answer in S use the command `dhyper(x, m, n, k)`:

```
> dhyper(25, 147, 3, 25)
[1] 0.576365
```

■

4.3 Continuous Univariate Distributions

4.3.1 Uniform Distribution (Continuous)

X is a **uniform** random variable defined on the interval $[a, b]$ if its **pdf** is given by

$$f(x|a, b) = \frac{1}{b-a}, \quad a \leq x \leq b.$$

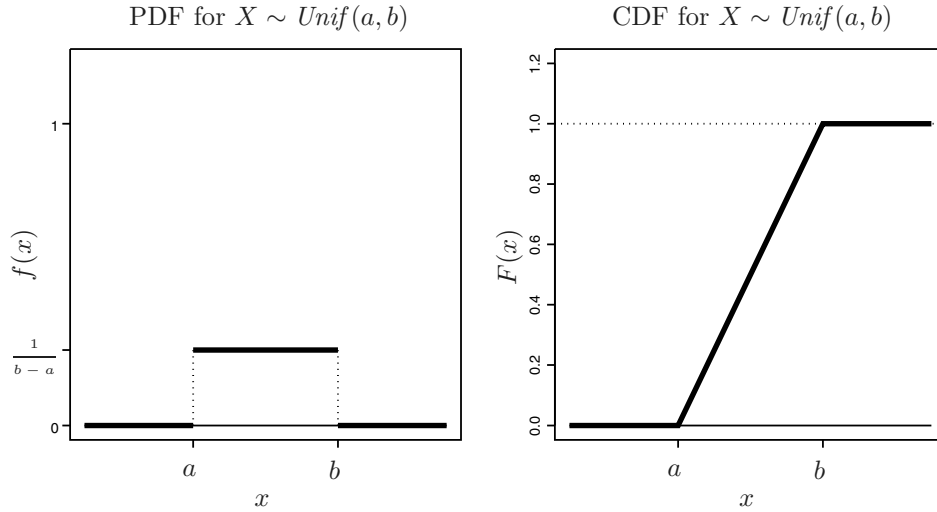
Some common uses of the uniform distribution include random number generation and modeling waiting times. The **pdf**, mean, variance, and **mgf** for a uniform random variable are found in (4.9).

<p>Uniform Distribution $X \sim \text{Unif}(a, b)$</p> $f(x a, b) = \frac{1}{b-a}, \quad a \leq x \leq b$ $E[X] = \frac{b+a}{2}$ $\text{Var}[X] = \frac{(b-a)^2}{12}$ $M_X(t) = \begin{cases} \frac{e^{tb} - e^{ta}}{t(b-a)} & \text{if } t \neq 0 \\ 1 & \text{if } t = 0 \end{cases}$	(4.9)
--	-------

Figure 4.4 on the facing page displays both the **pdf** and **cdf** for a $\text{Unif}(a, b)$ random variable.

Note that the area beneath the **pdf** is clearly one since the **pdf** forms a rectangle whose area is height \times length, $\frac{1}{(b-a)} \times (b-a) = 1$.

Example 4.12 Given a continuous random variable X defined over $[a, b]$ with **pdf** $f(x|a, b) = \frac{1}{b-a}$, $a \leq x \leq b$, find the expected value and the variance of X .

FIGURE 4.4: The **pdf** and **cdf** for the random variable $X \sim Unif(a, b)$

Solution: Using the definition for a continuous random variable from (3.12), write

$$\begin{aligned} E[X] &= \int_a^b x \cdot f(x) dx = \int_a^b \frac{x}{b-a} dx = \frac{1}{b-a} \cdot \frac{x^2}{2} \Big|_a^b \\ &= \frac{b^2 - a^2}{2(b-a)} = \frac{(b+a)(b-a)}{2(b-a)} = \frac{b+a}{2}. \end{aligned}$$

Next find $E[X^2]$ to use in computing the variance since $Var[X] = E[(X - \mu)^2] = E[X^2] - (E[X])^2$:

$$E[X^2] = \int_a^b x^2 \cdot \frac{1}{b-a} dx = \frac{1}{b-a} \cdot \frac{x^3}{3} \Big|_a^b = \frac{b^3 - a^3}{3(b-a)}$$

$$\begin{aligned} Var[X] &= E[X^2] - (E[X])^2 = \frac{b^3 - a^3}{3(b-a)} - \frac{(b+a)^2}{4} \\ &= \frac{(b-a)(b^2 + ab + a^2)}{3(b-a)} - \frac{(b+a)^2}{4} = \frac{4(b^2 + ab + a^2)}{12} - \frac{3(b+a)^2}{12} \\ &= \frac{4b^2 + 4ab + 4a^2 - (3b^2 + 6ab + 3a^2)}{12} = \frac{b^2 - 2ab + a^2}{12} \\ &= \frac{(b-a)^2}{12}. \end{aligned}$$

■

Example 4.13 If aerosol particles produced over forested areas have uniformly distributed diameters between 3 and 5 nanometers, compute the average volume of aerosol particles found over forested areas.

Solution: Recall that the volume of a sphere is $\frac{4}{3}\pi r^3$, or expressed in terms of the diameter, $\frac{1}{6}\pi d^3$. Consequently,

$$E\left[\frac{1}{6}\pi d^3\right] = \frac{1}{6}\pi E[d^3] \quad (4.10)$$

needs to be found. Let d represent the diameter of aerosol particles produced over forested areas. Since $d \sim Unif(3, 5)$,

$$\begin{aligned} E[d^3] &= \int_3^5 \frac{1}{5-3} \cdot x^3 dx = \frac{1}{2} \cdot \frac{x^4}{4} \Big|_3^5 \\ &= \frac{(5)^4}{8} - \frac{(3)^4}{8} = 68. \end{aligned}$$

Using the right side of (4.10), compute the average volume of aerosol particles to be

$$\frac{\pi}{6} \cdot 68 = 35.60472 \text{ nanometers}^3.$$

Estimate $E[d^3]$, denoted by $\widehat{E[d^3]}$, by cubing a large number of values drawn at random from a $Unif(3, 5)$ distribution and subsequently computing the mean of the cubed values. Then, the estimated mean volume of aerosol particles is computed by substituting $\widehat{E[d^3]}$ for $E[d^3]$ in the right-hand side of (4.10). The following S code estimates the mean volume of aerosol particles by simulating a sample of size 1000 from a $Unif(3, 5)$ distribution:

```
> (pi/6)*mean(runif(1000,3,5)^3)
[1] 35.61885
```

The simulated solution is within 0.02 of the theoretical solution. ■

Generating Pseudo-Random Numbers The generation of pseudo-random numbers is fundamental to any simulation study. The term “pseudo-random” is used because once one value in such a simulation is known, the next values can be determined without fail, since they are generated by an algorithm. Most major statistical software systems have reputable pseudo-random number generators. When using R, the user can specify one of several different random number generators, including a user-supplied random number generator. For more details, type `?RNG` at the R prompt. Generation of random values from named distributions is accomplished with the S command `rdist`, where `dist` is the distribution name; however, it is helpful to understand some of the basic ideas of random number generation in the event a simulation does not involve a named distribution. When the user wants to generate a sample from a continuous random variable X with **cdf** F , one approach is to use the *Inverse Transformation Method*. This method simply sets $F_X(X) = U \sim Unif(0, 1)$ and solves for X , assuming $F_X^{-1}(U)$ actually exists.

Example 4.14 Generate a sample of 1000 random values from a continuous distribution with **pdf** $f(x) = \frac{4}{3}x(2 - x^2)$, $0 \leq x \leq 1$. Verify that the mean and variance of the 1000 random values are approximately equal to the mean and variance of the given **pdf**.

Solution: First, the **cdf** is found. Then, $F_X(x)$ is set equal to u and solved.

$$F_X(x) = \int_0^x \frac{4}{3}t(2 - t^2) dt = \frac{4}{3} \left(x^2 - \frac{x^4}{4} \right) = \frac{1}{3}x^2(4 - x^2), \quad 0 \leq x \leq 1$$

Solving for x in terms of u by setting $u = F_X(x)$:

$$\begin{aligned}
 u &= \frac{1}{3}x^2(4-x^2) \\
 3u &= 4x^2 - x^4 && \text{multiply by 3 and distribute } x^2 \\
 -3u + 4 &= x^4 - 4x^2 + 4 && \text{multiply by } -1 \text{ and add 4 to complete the square} \\
 -3u + 4 &= (x^2 - 2)^2 && \text{factor} \\
 \pm\sqrt{-3u+4} &= x^2 - 2 && \text{take the square root of both sides} \\
 2 \pm \sqrt{-3u+4} &= x^2 && \text{add 2} \\
 \pm\sqrt{2 \pm \sqrt{-3u+4}} &= x && \text{take the square root of both sides,}
 \end{aligned}$$

which gives four solutions for x . The only one that is viable is $x = \sqrt{2 - \sqrt{4 - 3u}}$ because $0 \leq x \leq 1$. Provided $U \sim \text{Unif}(0, 1)$, $F_X^{-1}(U) = \sqrt{2 - \sqrt{4 - 3U}}$.

The theoretical mean and variance of X are calculated as

$$\begin{aligned}
 \mu_X = E(X) &= \int_0^1 x \cdot \frac{4}{3}x(2-x^2)dx = \frac{84}{135} = 0.6222222 \\
 E(X^2) &= \int_0^1 x^2 \cdot \frac{4}{3}x(2-x^2)dx = \frac{4}{9} = 0.4444444 \\
 \sigma_X^2 &= E(X^2) - E(X)^2 = \frac{4}{9} - \left(\frac{84}{135}\right)^2 = \frac{116}{2025} = 0.05728395
 \end{aligned}$$

The mean and variance of the 1000 simulated random values using `set.seed(33)` are 0.6152578 and 0.05809062, respectively, which are both within 2% of their theoretical values.

```

> set.seed(33)
> U <- runif(1000)
> X <- sqrt((2-sqrt(4-3*U)))
> mean(X)
[1] 0.6152578
> var(X)
[1] 0.05809062

```

Using numerical integration:

```

> f <- function(x){(4/3)*x*(2-x^2)}
> ex <- function(x){x*f(x)}
> ex2 <- function(x){x^2*f(x)}
> EX <- integrate(ex,0,1)
> EX2 <- integrate(ex2,0,1)
> VX <- EX2$value - EX$value^2
> c(EX$value, EX2$value, VX)
[1] 0.62222222 0.44444444 0.05728395

```

4.3.2 Exponential Distribution

When observing a Poisson process such as that in Example 4.4 on page 122, where the number of outcomes in a fixed interval such as the number of goals scored during 90 minutes of World Cup soccer is counted, the random variable X , which measures the number

of outcomes (number of goals), is modeled with the Poisson distribution. However, not only is X , the number of outcomes in a fixed interval, a random variable, but also is the waiting time between successive outcomes. If W is the waiting time until the first outcome of a Poisson process with mean $\lambda > 0$, then the **pdf** for W is

$$f(w) = \begin{cases} \lambda e^{-\lambda w} & \text{if } w \geq 0 \\ 0 & \text{if } w < 0 \end{cases}$$

Proof: Since waiting time is non-negative, $F(w) = 0$ for $w < 0$. When $w \geq 0$,

$$\begin{aligned} F(w) &= \mathbb{P}(W \leq w) = 1 - \mathbb{P}(W > w) \\ &= 1 - \mathbb{P}(\text{no outcomes in } [0, w]) \\ &= 1 - \frac{(\lambda w)^0 e^{-\lambda w}}{0!} \\ &= 1 - e^{-\lambda w} \end{aligned}$$

Consequently, when $w > 0$, the **pdf** of W is $F'(w) = f(w) = \lambda e^{-\lambda w}$.

The exponential distribution is characterized by a lack of memory property and is often used to model lifetimes of electronic components as well as waiting times for Poisson processes. A random variable is said to be **memoryless** if

$$\mathbb{P}(X > t_2 + t_1 | X > t_1) = \mathbb{P}(X > t_2) \text{ for all } t_1, t_2 \geq 0. \quad (4.11)$$

The **pdf**, mean, variance, and **mgf** for an exponential random variable are in (4.12), while the **pdf** and **cdf** for an exponential random variable are illustrated in Figure 4.5 on the facing page. The **cdf**, $F(x)$, for the exponential distribution is written

$$F(x) = \mathbb{P}(X \leq x) = \int_0^x \lambda e^{-\lambda t} dt = -e^{-\lambda t} \Big|_0^x = 1 - e^{-\lambda x}, \quad x \geq 0.$$

Exponential Distribution

$$X \sim \text{Exp}(\lambda)$$

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

$$E[X] = \frac{1}{\lambda}$$

$$\text{Var}[X] = \frac{1}{\lambda^2}$$

$$M_X(t) = (1 - \lambda^{-1}t)^{-1} \text{ for } t < \lambda$$

(4.12)

Example 4.15 Show that the function $f(x)$ in (4.12) satisfies condition 2 on page 93 from the properties of all **pdfs**.

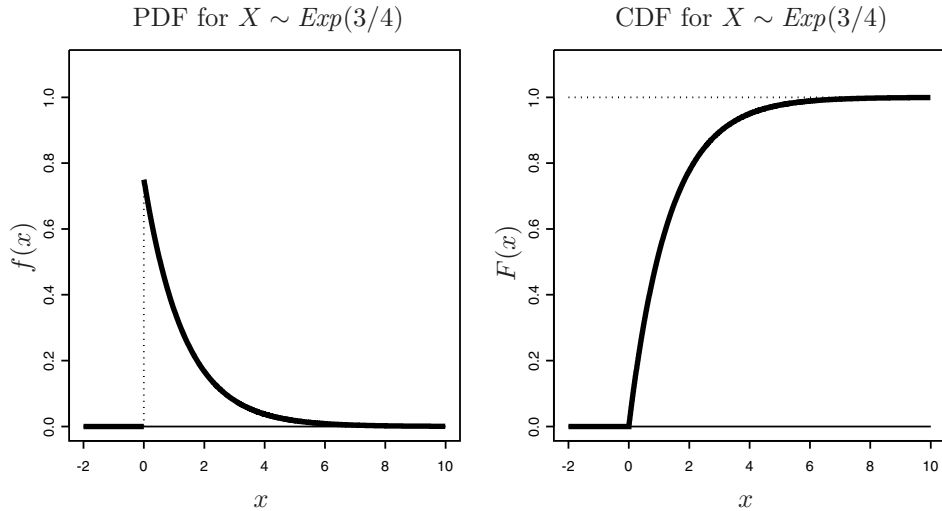


FIGURE 4.5: The **pdf** and **cdf** for the random variable $X \sim \text{Exp}(\lambda = 0.75)$

Solution: To satisfy condition 2 on page 93, it must be shown that the integral from $-\infty$ to $+\infty$ of the function $f(x)$ given in (4.12) is 1:

$$\begin{aligned} \int_{-\infty}^{\infty} \lambda e^{-\lambda x} dx &= \int_{-\infty}^0 0 dx + \int_0^{\infty} \lambda e^{-\lambda x} dx \\ &= \int_0^{\infty} \lambda e^{-\lambda x} dx = -e^{-\lambda x} \Big|_0^{\infty} = 0 - (-1) = 1. \end{aligned}$$

Example 4.16 Given $X \sim \text{Exp}(\lambda)$, find the mean and variance of X .

Solution: Using (3.12), write

$$E[X] = \int_0^{\infty} x \lambda e^{-\lambda x} dx.$$

Integrating by parts where $u = x$ and $dv = \lambda e^{-\lambda x} dx$, obtain

$$\begin{aligned} E[X] &= -xe^{-\lambda x} \Big|_0^{\infty} - \int_0^{\infty} -e^{-\lambda x} dx \\ &= 0 - \frac{1}{\lambda e^{\lambda x}} \Big|_0^{\infty} = \frac{1}{\lambda}. \end{aligned}$$

Before finding the variance of X , find $E[X^2]$ using (3.13) as follows:

$$E[X^2] = \int_0^{\infty} x^2 \lambda e^{-\lambda x} dx \tag{4.13}$$

Note that $E[X] = \int_0^{\infty} x\lambda e^{-\lambda x} dx \Rightarrow \frac{E[X]}{\lambda} = \int_0^{\infty} x e^{-\lambda x} dx$ and integrate (4.13) by parts where $u = x^2$ and $dv = \lambda e^{-\lambda x} dx$:

$$\begin{aligned} E[X^2] &= -x^2 e^{-\lambda x} \Big|_0^{\infty} - \int_0^{\infty} -2x e^{-\lambda x} dx \\ &= 0 + 2 \frac{E[X]}{\lambda} = \frac{2}{\lambda^2}. \end{aligned}$$

Using the fact that $\text{Var}[X] = E[X^2] - (E[X])^2$, obtain $\text{Var}[X] = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$. ■

Based on the results from Example 4.16, note that the mean and standard deviation of the exponential random variable are identical. Quite often, the **pdf** for the exponential is expressed as

$$f(x) = \frac{1}{\theta} e^{-x/\theta}, \quad x \geq 0, \quad \theta > 0,$$

where $\theta = \frac{1}{\lambda}$. Of course, the **mgf** is then written as $M_X(t) = (1 - \theta t)^{-1}$ and the reparameterized mean and variance are θ and θ^2 , respectively. Note the relationship between the Poisson mean and the exponential mean. Given a Poisson process with mean λ , the waiting time until the first outcome has an exponential distribution with mean $\frac{1}{\lambda}$. That is, if λ represents the number of outcomes in a unit interval, $\frac{1}{\lambda}$ is the mean waiting time for the first change. If X denotes the lifetime of an electronic component following an exponential distribution with mean $\frac{1}{\lambda}$, (4.11) implies that the probability the component will work for $t_2 + t_1$ hours given that it has worked for t_1 hours is the same as the probability that the component will function for at least t_2 hours. In other words, the component has no memory of having functioned for t_1 hours. Note that (4.11) is equivalent to

$$\frac{\mathbb{P}(X > t_2 + t_1, X > t_1)}{\mathbb{P}(X > t_1)} = \mathbb{P}(X > t_2),$$

which is equivalent to

$$\mathbb{P}(X > t_2 + t_1) = \mathbb{P}(X > t_2)\mathbb{P}(X > t_1). \quad (4.14)$$

Since $\mathbb{P}(X > t_2 + t_1) = e^{-\lambda(t_2+t_1)} = e^{-\lambda t_2} e^{-\lambda t_1} = \mathbb{P}(X > t_2)\mathbb{P}(X > t_1)$ for any exponential random variable, exponential random variables are memoryless according to (4.14).

Example 4.17 ▷ **Exponential Distribution: Light Bulbs** ◁ If the life of a certain type of light bulb has an exponential distribution with a mean of 8 months, find

- (a) The probability that a randomly selected light bulb lasts between 3 and 12 months.
- (b) The 95th percentile of the distribution.
- (c) The probability that a light bulb that has lasted for 10 months will last more than 25 months.

Solution: The answers are as follows:

(a) Since $X \sim \text{Exp}(\lambda = \frac{1}{8})$, the probability that a randomly selected light bulb lasts between 3 and 12 months is

$$\mathbb{P}(3 < X < 12) = \int_3^{12} \frac{1}{8} e^{-x/8} dx = -e^{-x/8} \Big|_3^{12} = -0.2231 + 0.6873 = 0.4642.$$

The following code solves the problem with S:

```
> round(pexp(12,1/8) - pexp(3,1/8),4)
[1] 0.4642
```

The function `integrate()` can also be used to solve this problem using numerical integration:

```
> f1 <- function(x){(1/8)*exp(-x/8)}
> integrate(f1,3,12) # For R
0.4641591 with absolute error < 5.2e-15

> f1 <- function(x){(1/8)*exp(-x/8)}
> round(integrate(f1,3,12)$integral,4) # For S-PLUS
[1] 0.4642
```

(b) The 95th percentile is the value x_{95} such that

$$\begin{aligned} \int_{-\infty}^{x_{95}} f(x) dx &= \int_0^{x_{95}} \frac{1}{8} e^{-x/8} dx = \frac{95}{100} \\ -e^{-x/8} \Big|_0^{x_{95}} &= 1 - e^{-\frac{x_{95}}{8}} = \frac{95}{100} \\ e^{-\frac{x_{95}}{8}} &= \frac{5}{100} \\ x_{95} &= -8 \ln(0.05) = 23.96586 \end{aligned}$$

To find the answer with S, type

```
> qexp(0.95,1/8)
[1] 23.96586
```

(c) The probability that a light bulb that has lasted for 10 months will last more than 25 months mathematically is written $\mathbb{P}(X > 25|X > 10)$. Because an exponential distribution is present, (4.11) can be used to say that this is equal to $\mathbb{P}(X > 15) = e^{-15/8} = 0.153355$.

Solve the problem with S as follows:

```
> 1-pexp(15,1/8)
[1] 0.1533550
```

Example 4.18 ▷ *Exponential Distribution: Intergoal Times* ◁ Example 4.4 on page 122 illustrated how the number of goals scored during World Cup games could be modeled with the Poisson distribution. Now, look at the distribution of T , the time between goals. In Example 4.4 on page 122, λ was estimated to be $\frac{575}{232}$. Since one soccer match lasts 90 minutes, the average time (in minutes) before a goal is scored is $\frac{90}{\lambda} = 36.31304$ minutes assuming λ is $\frac{575}{232}$. To find the intergoal times from the cumulative goal times stored in column `CGT` of the `Soccer` data frame, compute `CGTi+1 - CGTi`. ■

- Compute the mean and standard deviation for the time between goals.
- Is it reasonable to model the time between goals with the exponential distribution?
- In particular, is the lack of memory property evident in the data?

Solution: The answers are as follows:

(a) First, attach `Soccer` so that columns can be referenced by their names. Then, use `S` to calculate both the mean and standard deviation for the time between goals:

```
> attach(Soccer)
> inter.times <- CGT[2:575] - CGT[1:574]
> mean(inter.times)
[1] 36.24042
> sd(inter.times)
[1] 36.67138
```

Note that both the mean and standard deviation for time between goals are close to the theoretical time of 36.31 minutes under the assumption that λ is $\frac{575}{232}$.

(b) To assess the fit of the data to an exponential distribution with a mean of 36.31 minutes, first split the data into discrete categories. If the underlying distribution is exponential, then a good bin width is approximately $(\frac{12}{n})^{1/3} \cdot \mu_X$ (Scott, 1992). In our case, the bin width is $(\frac{12}{574})^{1/3} \cdot 36.31 \approx 10$.

```
> rate <- 1/(90/(575/232))
> ntot <- length(inter.times)
> OBS <- table(cut(inter.times, breaks=c(seq(0,130,10), 330)))
> EmpiP <- round(OBS/ntot,3)
> TheoP <- round(c((pexp(seq(10,130,10),rate) - pexp(seq(0,120,10),rate)),
+ (1 - pexp(130, rate))), 3)
> EXP <-round(TheoP*ntot, 0)
> ANS <-cbind(OBS, EXP, EmpiP, TheoP)
> ANS
```

	OBS	EXP	EmpiP	TheoP
(0,10]	144	138	0.251	0.241
(10,20]	106	105	0.185	0.183
(20,30]	86	80	0.150	0.139
(30,40]	53	60	0.092	0.105
(40,50]	45	46	0.078	0.080
(50,60]	27	35	0.047	0.061
(60,70]	35	26	0.061	0.046
(70,80]	16	20	0.028	0.035
(80,90]	22	15	0.038	0.027
(90,100]	12	11	0.021	0.020
(100,110]	3	9	0.005	0.015
(110,120]	3	7	0.005	0.012
(120,130]	6	5	0.010	0.009
(130,330]	16	16	0.028	0.028

The observed and expected values as well as the empirical and theoretical probabilities are similar.

(c) The lack of memory property is also evident from the data. Empirically, $\mathbb{P}(T > 10) = 1 - \mathbb{P}(T \leq 10) = 1 - \frac{144}{574} = \frac{430}{574} = 0.749$, and $\mathbb{P}(T > 20 | T > 10) = \frac{574-144-106}{574-144} = 0.754$, which are both roughly the same and similar to the theoretical $\mathbb{P}(T > 10)$, which is 0.759 under the assumption that the mean is 36.31 minutes. Since the observed data appear to lack memory, the same probability statements could be used to justify independence among the times between goals using (4.14). Finally, produce a histogram of the observed data

similar to Figure 4.6, and superimpose over this the density for an exponential with a mean of 36.31 minutes. Based on the analysis and Figure 4.6, it seems reasonable to model the time between goals scored in World Cup competition for the years 1990 to 2002 with an exponential distribution:

```
> hist(inter.times, breaks=seq(0,310,10), col=13, xlim=c(0,125), prob=TRUE,
+ xlab="Time Between Goals")
> xt <- seq(0,140,0.01)
> ft <- dexp(xt, rate)
> lines(xt, ft, type="l")
> detach(Soccer) ■
```

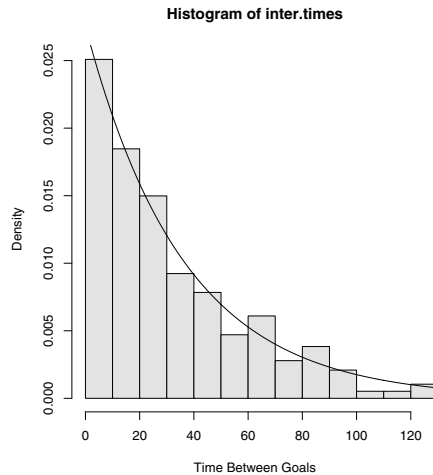


FIGURE 4.6: Histogram of time between goals with superimposed exponential density curve with mean of 36.31 minutes

4.3.3 Gamma Distribution

Some random variables are always non-negative and yield distributions of data that tend to be skewed. The waiting time until a certain number of malfunctions in jet engines, the waiting time until a certain number of accidents at a given intersection, and similar scenarios where the random variable of interest is the waiting time until a certain number of events takes place yield skewed distributions. The **gamma** distribution is often used to model the waiting time until the α^{th} event in a Poisson process. Before defining the gamma distribution, review the definition of the gamma function from mathematics. The **gamma function**, $\Gamma(\alpha)$, is defined by:

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx, \quad \alpha > 0 \quad (4.15)$$

Some of the more important properties of the gamma function include:

1. For $\alpha > 0$, $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$.
2. For any positive integer, n , $\Gamma(n) = (n - 1)!$
3. $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

In Section 4.3.2 on page 133, it was proved that the waiting time until the first outcome in a Poisson process follows an exponential distribution. Now, let W denote the waiting time until the α^{th} outcome and derive the distribution of W in a similar fashion. Since waiting time is non-negative, $F(w) = 0$ for $w < 0$. When $w \geq 0$,

$$\begin{aligned} F(w) &= \mathbb{P}(W \leq w) = 1 - \mathbb{P}(W > w) \\ &= 1 - \mathbb{P}(\text{fewer than } \alpha \text{ outcomes in } [0, w]) \\ &= 1 - \sum_{k=0}^{\alpha-1} \frac{(\lambda w)^k e^{-\lambda w}}{k!} \end{aligned}$$

Consequently, when $w > 0$, the **pdf** of W is $F'(w) = f(w)$ whenever this derivative exists. It follows then that

$$\begin{aligned} f(w) = F'(w) &= - \sum_{k=0}^{\alpha-1} \frac{(\lambda w)^k e^{-\lambda w} (-\lambda) + e^{-\lambda w} k(\lambda w)^{k-1} \lambda}{k!} \\ &= -e^{-\lambda w} \sum_{k=0}^{\alpha-1} \frac{k\lambda(\lambda w)^{k-1} - \lambda(\lambda w)^k}{k!} \\ &= \lambda e^{-\lambda w} - e^{-\lambda w} \sum_{k=1}^{\alpha-1} \frac{k\lambda(\lambda w)^{k-1} - \lambda(\lambda w)^k}{k!} \\ &= \lambda e^{-\lambda w} - e^{-\lambda w} \sum_{k=1}^{\alpha-1} \left[\frac{\lambda(\lambda w)^{k-1}}{(k-1)!} - \frac{\lambda(\lambda w)^k}{k!} \right] \\ &= \lambda e^{-\lambda w} - e^{-\lambda w} \left[\frac{\lambda(\lambda w)^0}{0!} - \frac{\lambda(\lambda w)^1}{1!} + \frac{\lambda(\lambda w)^1}{1!} - \frac{\lambda(\lambda w)^2}{2!} + \right. \\ &\quad \left. \dots - \frac{\lambda(\lambda w)^{\alpha-1}}{(\alpha-1)!} \right] \\ &= \lambda e^{-\lambda w} - e^{-\lambda w} \left[\lambda - \frac{\lambda(\lambda w)^{\alpha-1}}{(\alpha-1)!} \right] \\ &= \frac{\lambda(\lambda w)^{\alpha-1} e^{-\lambda w}}{(\alpha-1)!} = \frac{\lambda^\alpha w^{\alpha-1} e^{-\lambda w}}{\Gamma(\alpha)} \end{aligned}$$

From the previous derivation, note that the gamma is a generalization of the exponential distribution. The **pdf**, mean, variance, and **mgf** for a gamma random variable are listed in (4.16). The **pdfs** for $\lambda = 2$ and $\lambda = 1$ with $\alpha = 1, 2$, and 3 , respectively, are illustrated in Figure 4.7 on the facing page. Notice that different shapes are produced in Figure 4.7 for different values of α . For this reason, α is often called the shape parameter associated with the gamma distribution. The parameter λ is referred to as the scale parameter. Varying λ changes the units of measurement (say, from seconds to minutes) and does not affect the shape of the density.

Gamma Distribution

$X \sim \Gamma(\alpha, \lambda)$

$$f(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

$$E[X] = \frac{\alpha}{\lambda}$$

$$\text{Var}[X] = \frac{\alpha}{\lambda^2}$$

$$M_X(t) = (1 - \lambda^{-1}t)^{-\alpha} \text{ for } t < \lambda$$

(4.16)

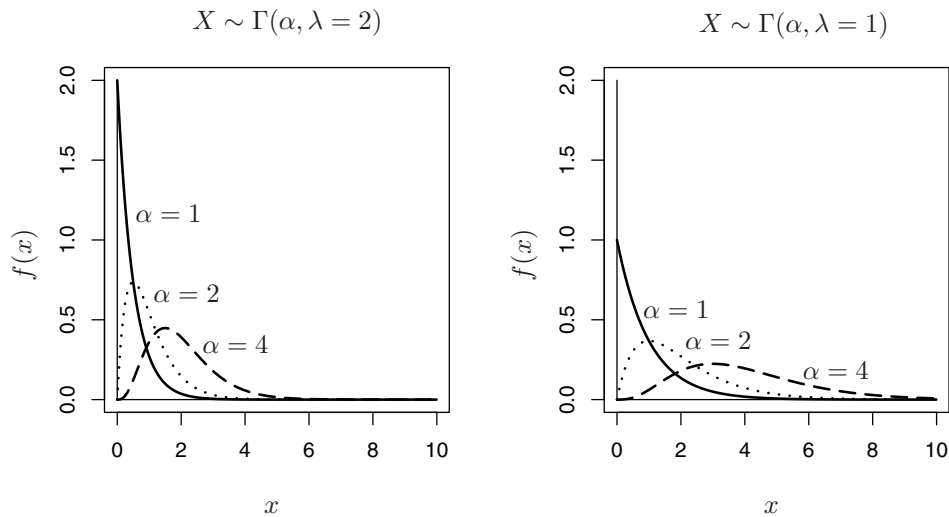


FIGURE 4.7: Graphical illustration of the **pdfs** of a $\Gamma(\alpha, 2)$ and a $\Gamma(\alpha, 1)$ random variable for $\alpha = 1, 2,$ and $4,$ respectively.

Useful Relationships

1. Given $X \sim \Gamma(\alpha, \lambda)$. When $\alpha = 1$, the resulting random variable is $X \sim \text{Exp}(\lambda)$. That is, the exponential distribution is a special case of the gamma distribution.
2. Given $X \sim \Gamma(\alpha, \lambda)$. When $\alpha = n/2$ and $\lambda = 1/2$, the resulting random variable has a chi-square distribution with n degrees of freedom. (The chi-square is discussed in Section 6.6.1.)
3. Given $X \sim \Gamma(\alpha, \lambda)$. Provided α is a positive integer, the resulting distribution is known as the Erlang. In this case, the Erlang distribution gives the waiting time until the α^{th} occurrence when the number of outcomes in an interval of length t follows a Poisson distribution with parameter λt .

Example 4.19 Given $X \sim \Gamma(\alpha, \lambda)$, find the mean and variance of X .

Solution: Using the **mgf** from (4.16), it is known that the first and second derivatives of the **mgf** evaluated at zero, respectively, yield the $E[X]$ and the $E[X^2]$. Consequently,

$$\begin{aligned} E[X] &= M'_X(t) \Big|_{t=0} \\ &= (-\alpha) (1 - \lambda^{-1}t)^{-\alpha-1} (-\lambda^{-1}) \Big|_{t=0} = \frac{\alpha}{\lambda} \\ E[X^2] &= M''_X(t) \Big|_{t=0} \\ &= \alpha\lambda^{-1}(-\alpha-1) (1 - \lambda^{-1}t)^{-\alpha-2} (-\lambda^{-1}) \Big|_{t=0} = \frac{\alpha(\alpha+1)}{\lambda^2} \\ \text{Var}[X] &= E[X^2] - (E[X])^2 \\ &= \frac{\alpha(\alpha+1)}{\lambda^2} - \left(\frac{\alpha}{\lambda}\right)^2 = \frac{\alpha}{\lambda^2} \end{aligned}$$

So the mean of X is $\frac{\alpha}{\lambda}$ and the variance of X is $\frac{\alpha}{\lambda^2}$. ■

Example 4.20 Suppose that the average arrival rate at a local fast food drive-through window is three cars per minute ($\lambda = 3$). Find

- The probability that at least five cars arrive in 120 seconds.
- The probability that more than one minute elapses before the second car arrives.
- If one car has already gone through the drive-through, what is the average waiting time before the third car arrives?

Solution: The answers are as follows:

(a) If the average number of car arrivals follows a Poisson distribution with a rate of three cars per minute, then the average rate of arrival for 2 minutes is six cars. Given that $X \sim \text{Pois}(\lambda = 6)$,

$$\mathbb{P}(X \geq 5) = 1 - \mathbb{P}(X \leq 4) = 1 - \sum_{x=0}^4 \frac{e^{-6}6^x}{x!} = 1 - 0.2850565 = 0.7149435.$$

To solve the problem with S, use the command `ppois()`:

```
> 1 - ppois(4,6)
[1] 0.7149435
```

(b) Let W represent the waiting time until the α^{th} outcome. It follows that $W \sim \Gamma(\alpha = 2, \lambda = 3)$. Consequently,

$$\begin{aligned} \mathbb{P}(W > 1) &= 1 - \mathbb{P}(W \leq 1) = 1 - \mathbb{P}(\Gamma(2, 3) \leq 1) = 1 - \int_0^1 \frac{3^2}{\Gamma(2)} x^{2-1} e^{-3x} dx \\ &= 1 - \int_0^1 3x e^{-3x} dx \end{aligned}$$

Using integration by parts where $u = 3x$ and $dv = 3e^{-3x} dx$,

$$\begin{aligned} \int_0^1 3x e^{-3x} 3 dx &= -3xe^{-3x} \Big|_0^1 + \int_0^1 3e^{-3x} dx \\ &= -3e^{-3} + \left[-e^{-3x} \Big|_0^1 \right] = -3e^{-3} + \left[-e^{-3} + 1 \right] \\ &= 1 - 4e^{-3} = 0.8008517. \end{aligned}$$

In other words, $\mathbb{P}(W > 1) = 1 - 0.8008517 = 0.1991483$. To solve the problem with **S**, use the command `pgamma()` or `integrate()`:

```
> 1 - pgamma(1,2,3)
[1] 0.1991483
```

```
> gam23<-function(x){9*x*exp(-3*x)}
> integrate(gam23,1, Inf) # R
0.1991483 with absolute error < 2.5e-05
```

```
> gam23<-function(x){9*x*exp(-3*x)}
> integrate(gam23,1, Inf)$integral # S-PLUS
[1] 0.1991483
```

(c) This problem is really asking for the mean of a $\Gamma(\alpha = 2, \lambda = 3)$ random variable. Note: $\alpha = 2$ since one car has already arrived and the problem requests the average waiting time until the third car arrives. Therefore, $E[X] = \frac{\alpha}{\lambda} = \frac{2}{3}$. In other words, there is an average wait of $\frac{2}{3}$ of a minute before the arrival of the third vehicle given one vehicle has already arrived. ■

4.3.4 Hazard Function, Reliability Function, and Failure Rate

In addition to studying the **pdf** of continuous random variables, at times it is helpful to study other functions related to the **pdf** such as the **reliability function** or the **hazard function** which is also often called the **failure rate** or **force of mortality**, especially when dealing with lifetime data. Suppose the random variable T represents the useful life of some component with **pdf** and **cdf** given by $f(t)$ and $F(t)$, respectively. The reliability function $R(t)$ is defined as

$$R(t) = \mathbb{P}(T > t) = 1 - F(t), \quad t > 0 \quad (4.17)$$

and represents the probability that the lifetime of the component exceeds t . The hazard function, $h(t)$, is defined as

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{R(t)}, \quad t > 0, \quad F(t) < 1. \quad (4.18)$$

Note that the hazard function is often called the conditional failure rate.

The functions $h(t)$, $f(t)$, and $F(t)$ provide mathematically equivalent specifications of the distribution of T . In fact, it can be shown that

$$f(t) = h(t)e^{-\int_0^t h(x) dx}. \quad (4.19)$$

To gain an intuitive understanding of what $h(t)$ is measuring, let dt represent a small unit of measurement. Then, the quantity $h(t)dt$ can be thought of as the approximate probability

that T takes on a value in $(t, t + dt)$. Keeping in mind that $1 - F(t) = \mathbb{P}(T > t)$, write

$$h(t)dt = \frac{f(t)dt}{1 - F(t)} \approx \mathbb{P}[T \in (t, t + dt) | T > t].$$

In other words, $h(t)dt$ represents the approximate probability of having a breakdown during the interval $(t, t + dt)$ given that a component has lasted up to time t . In mathematical terms,

$$\lim_{dt \rightarrow 0} \frac{\mathbb{P}(t \leq T \leq t + dt | T > t)}{dt} \quad (4.20)$$

may be written, which represents the instantaneous rate of death or failure at time t , given the individual or component has survived to time t . It may then be noted that the hazard function is a rate rather than a probability. The failure rate for an exponential random variable is a constant λ :

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{\lambda e^{-\lambda t}}{1 - [1 - e^{-\lambda t}]} = \lambda.$$

Not many components have a constant failure rate. As a matter of fact, it stands to reason that the failure rate should increase as the life of a component ages. For most manufactured items as well as human populations, this is the case after some initial time period. However, there are some instances such as breakdowns when equipment is on a preventative maintenance schedule where it is still reasonable to assume a constant failure rate. A very flexible hazard function is $h(t) = \frac{\alpha t^{\alpha-1}}{\beta^\alpha}$, for all α and β greater than 0, since the function is monotone increasing for $\alpha > 1$, monotone decreasing for $\alpha < 1$, and constant for $\alpha = 1$, as illustrated in Figure 4.8. This hazard function corresponds to the Weibull distribution that is discussed in Section 4.3.5

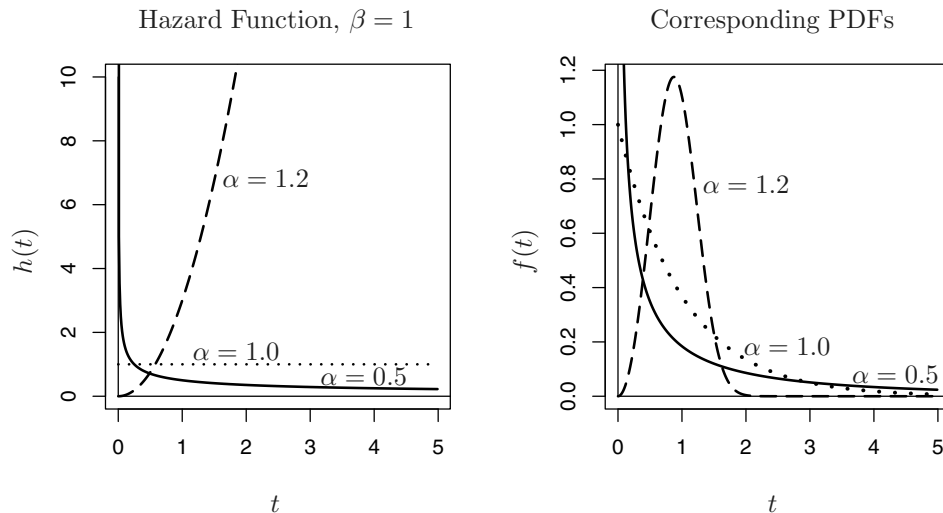


FIGURE 4.8: Illustration of the hazard function $h(t) = \frac{\alpha t^{\alpha-1}}{\beta^\alpha}$ for $\alpha = 0.5$, $\alpha = 1.0$, and $\alpha = 1.2$ with $\beta = 1$ and the corresponding pdfs

Example 4.21 ▷ **Hazard Rate** ◁ In an effort to attract more business, a local computer outlet has agreed to replace its laser printers with a brand new laser printer in the event any of its laser printers malfunction within one year of the date of their purchase. According to the manufacturer of the printer, the useful life (in years) of the printer is a random variable T with **pdf** $f(t) = K(2000 - 0.1e^{-2t})$ for $0 < t < 5$.

- Find K so that $f(t)$ is a **pdf**.
- Compute the probability a randomly selected laser printer will have to be replaced due to a malfunction.
- What are the mean and standard deviation for laser printer life?
- If a small business purchases five laser printers from the computer outlet, what is the probability there are no malfunctions during the first year?
- What should the length of guarantee time be for a laser printer if the outlet store wants to replace no more than 5% of the laser printers?
- Compute, graphically represent, and interpret the hazard function.

Solution: The answers are as follows:

- To find K such that $f(t)$ is a **pdf**, the integral over all possible values of t must be one:

$$\int_0^5 K(2000 - 0.1e^{-2t}) dt \stackrel{\text{set}}{=} 1$$

$$K \left[(2000t + 0.05e^{-2t}) \Big|_0^5 \right] = 1$$

$$K [10000 + 0.05e^{-10} - 0.05] = 1$$

$$K = \frac{1}{9999.95 + 0.05e^{-10}}$$

Let the denominator of K be equal to $k1 = 9999.95 + 0.05e^{-10}$ for the remainder of the problem. The solution given is for R. To obtain similar answers with S-PLUS, replace **\$value** with **\$integral** throughout. To calculate the denominator of K numerically with R, enter

```
> g <- function(x){(2000 - 0.1*exp(-2*x))}
> k1a <- integrate(g,0,5)$value
> k1a
[1] 9999.95
> # OR
> k1 <- (10000 +0.05*exp(-10) -0.05)
> k1
[1] 9999.95
> f <- function(x){1/k1*(2000 - 0.1*exp(-2*x))}
> integrate(f,0,5)
1 with absolute error < 1.1e-14
```

$$(b) P(T < 1) = \int_0^1 \frac{1}{k1} (2000 - 0.1e^{-2t}) dt = 0.1999967$$

With S using the f from part (a):

```
> integrate(f,0,1)
0.1999967 with absolute error < 2.2e-15
```

$$(c) E(T) = \int_0^5 \frac{1}{k1} t (2000 - 0.1e^{-2t}) dt = 2.50001$$

```
> et <- function(x){x*f(x)}
> ET <- integrate(et,0,5)$value
> ET
[1] 2.50001
```

$$\sigma_T = \sqrt{\sigma_T^2} = \sqrt{E(T^2) - E(T)^2} = 1.443372$$

```
> et2 <- function(x){x^2*f(x)}
> ET2 <- integrate(et2,0,5)$value
> VX <- ET2 - ET^2
> SX <- sqrt(VX)
> SX
[1] 1.443372
```

(d) Assuming the useful lives of laser printers are independent, the probability none of the five printers have to be replaced is

$$P(T_1 > 1) \times P(T_2 > 1) \times \cdots \times P(T_5 > 1) = (1 - 0.1999967)^5 = 0.3276868$$

If the random variable X is defined to be the number of printers that need to be replaced during the first year of operation, then $X \sim Bin(n = 5, \pi = 0.1999967)$ and the problem is solved by computing $P(X = 0) = \binom{5}{0} (0.1999967)^0 (1 - 0.1999967)^5 = 0.3276868$:

```
> dbinom(0,5,0.1999967)
[1] 0.3276868
```

(e) The length of guarantee time for a laser printer if the outlet store wants to replace no more than 5% of the laser printers will be the roots of the equation

$$\mathbb{P}(T < t) = \int_0^t \frac{1}{k1} (2000 - 0.1e^{-2x}) dx = 0.05.$$

$$\int_0^t \frac{1}{k1} (2000 - 0.1e^{-2x}) dx = 0.05$$

$$(2000x + 0.05e^{-2x}) \Big|_0^t = 0.05k1$$

$$2000t + 0.05e^{-2t} - 0.05 = 0.05k1$$

$$\text{Find roots of } 2000t + 0.05e^{-2t} - 0.05 - 0.05k1 = 0$$

Use the function `uniroot()` to solve for t numerically:

```
> fr <- function(x){2000*x+0.05*exp(-2*x)-0.05*k1 -0.05}
> uniroot(fr, c(0,5))$root
[1] 0.25
```

Since t is given in years, multiplying $0.25 \times 365 = 91.25$ days. In other words, the computer outlet will have to replace less than 5% of their laser printers if they use a guarantee period of 91 days.

(f) Note that the reliability (survival) function is

$$P(T > t) = 1 - F(t) = 1 - \frac{1}{k_1} (2000t + 0.05e^{-2t} - 0.05), \quad 0 < t < 5.$$

Using the reliability function, the hazard function can be written as

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{\frac{1}{k_1} (2000 - 0.1e^{-2t})}{1 - \frac{1}{k_1} (2000t + 0.05e^{-2t} - 0.05)}, \quad 0 < t < 5.$$

This particular hazard function (Figure 4.9) represents the instantaneous rate of failure given that a printer has lasted until time t . The R commands used to create Figure 4.9 follow. Note that `f()` (used for `f(year)`) was defined in part (a).

```
> year <- seq(0,5, length=500)
> CDF <- function(x){1/k1*(2000*x + 0.05*exp(-2*x)-0.05)}
> plot(year, f(year)/(1-CDF(year)), type="l", ylab="h(year)")
```

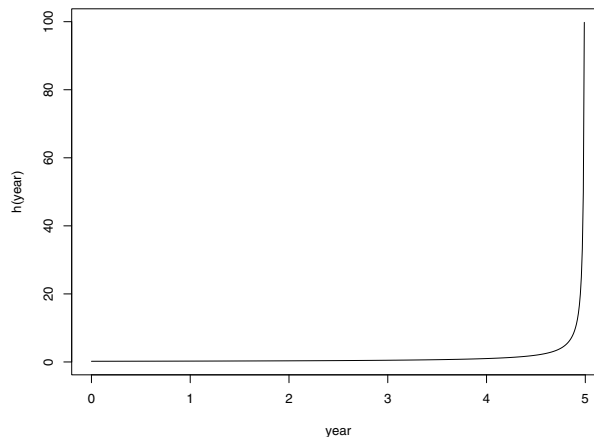


FIGURE 4.9: Hazard function for printer failure

■

4.3.5 Weibull Distribution

The gamma distribution provides an adequate model for some systems' lifetime distributions. However, since the hazard function for the gamma does not have a closed form expression, and its failure rate approaches λ from both above (when $\alpha < 1$) and below (when $\alpha > 1$) as t gets large, distributions with closed form expressions for the hazard function such as the Weibull tend to be favored by practitioners who deal with lifetime

distributions. In particular, the hazard function for the Weibull distribution has a failure rate that varies with time. The hazard rate for the Weibull distribution is $h(t) = \frac{\alpha t^{\alpha-1}}{\beta^\alpha}$, for all α and β greater than 0. Using (4.19), derive the **pdf** for the Weibull distribution as follows:

$$f(t) = h(t)e^{-\int_0^t h(x) dx} = \frac{\alpha t^{\alpha-1}}{\beta^\alpha} e^{-\int_0^t \frac{\alpha x^{\alpha-1}}{\beta^\alpha} dx} = \frac{\alpha t^{\alpha-1}}{\beta^\alpha} e^{-(t/\beta)^\alpha}$$

The **pdf**, mean, variance, and hazard function for a Weibull random variable ($\alpha > 0$ and $\beta > 0$) are in (4.21), while the **pdfs** for $Weib(\alpha, 1)$ and a $Weib(\alpha, 2)$ random variable for $\alpha = 1, 2,$ and $5,$ respectively, are illustrated in Figure 4.10. As with the gamma distribution, the first parameter in the Weibull distribution, $\alpha,$ is the shape parameter; and the second parameter, $\beta,$ is the scale parameter. If $X \sim Weib(\alpha, \beta)$ and $\alpha = 1,$ then $X \sim Exp(\lambda = 1/\beta).$

Weibull Distribution
 $X \sim Weib(\alpha, \beta)$

$$f(x) = \begin{cases} \alpha\beta^{-\alpha}x^{\alpha-1}e^{-(x/\beta)^\alpha} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \quad (4.21)$$

$$E[X] = \beta\Gamma(1 + \alpha^{-1})$$

$$Var[X] = \beta^2 \left\{ \Gamma(1 + 2\alpha^{-1}) - [\Gamma(1 + \alpha^{-1})]^2 \right\}$$

$$h(x) = \alpha\beta^{-\alpha}x^{\alpha-1} \text{ for } x \geq 0$$

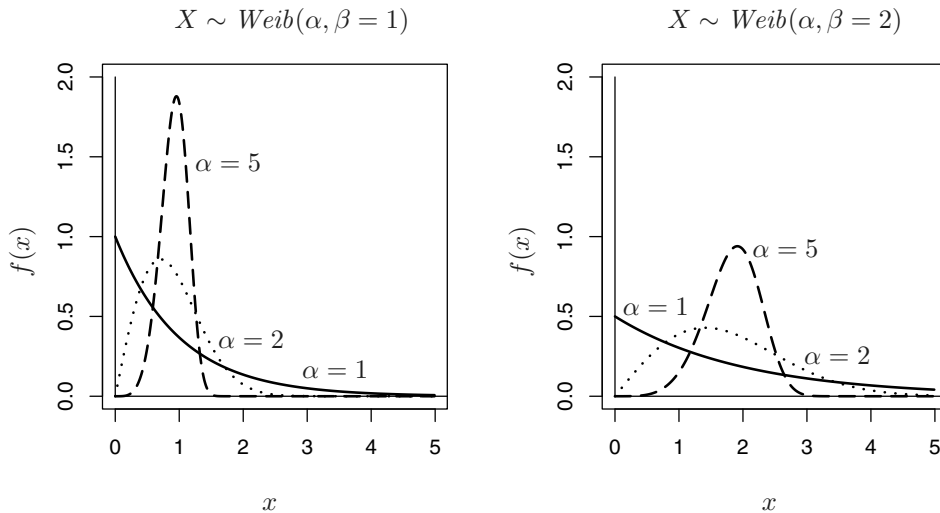


FIGURE 4.10: Illustration of the **pdfs** of a $Weib(\alpha, 1)$ and a $Weib(\alpha, 2)$ random variable for $\alpha = 1, 2,$ and $5,$ respectively

Example 4.22 The useful life (in thousands of hours) of a certain type of transistor follows a Weibull distribution with $\alpha = 2$ and $\beta = 8$. Find the probability that a randomly selected transistor lasts more than 8000 hours. What is the average life for this type of transistor?

Solution: First, find the **cdf** for $X \sim Weib(\alpha, \beta)$:

$$F(x) = \int_0^x \alpha \beta^{-\alpha} t^{\alpha-1} e^{-(t/\beta)^\alpha} dt = -e^{-(t/\beta)^\alpha} \Big|_0^x = 1 - e^{-(x/\beta)^\alpha}$$

Using the **cdf** for the Weibull, the probability a randomly selected transistor lasts more than 8000 hours is

$$\mathbb{P}(X > 8) = 1 - F(8) = 1 - \left[1 - e^{-(8/8)^2}\right] = e^{-1} = 0.3678794.$$

The expected value of X (in thousands of hours) is

$$E[X] = \beta \Gamma(1 + \alpha^{-1}) = 8 \Gamma\left(1 + \frac{1}{2}\right) = 8 \frac{1}{2} \Gamma\left(\frac{1}{2}\right) = 4\sqrt{\pi} = 7.089815.$$

To solve the first question and to compute $\Gamma\left(\frac{3}{2}\right)$ with **S**, use the functions `pweibull()` and `gamma()`, respectively:

```
> 1 - pweibull(8,2,8)
[1] 0.3678794
> 8*gamma(3/2)
[1] 7.089815
```

■

4.3.6 Beta Distribution

The continuous distributions discussed up to this point, with the exception of the continuous uniform, have positive densities over unbounded intervals. To model phenomena restricted to a finite interval, another type of distribution is needed, such as the beta (β) distribution, whose density function is positive only over the interval (A, B) . The standard beta distribution, ($A = 0, B = 1$), is often used to model proportions, especially in Bayesian analysis, where parameters are treated as random variables. For example, π from the binomial distribution can be modeled with the standard β distribution as it takes on only non-zero values in the interval $(0, 1)$. The distribution can take on a wide variety of shapes, as depicted in Figure 4.11 on the following page. The **pdf**, mean, and variance for a general beta random variable ($\alpha > 0$ and $\beta > 0$) are in (4.22). When working with the standard β distribution, that is, $A = 0$ and $B = 1$, a β random variable X is denoted simply $X \sim \beta(\alpha, \beta)$. The β distribution available in **S** is the standard β distribution rather than the general β distribution.

Beta Distribution
 $X \sim \beta(\alpha, \beta, A, B)$

$$f(x) = \begin{cases} \frac{1}{B-A} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \left(\frac{x-A}{B-A}\right)^{\alpha-1} \left(\frac{B-x}{B-A}\right)^{\beta-1} & \text{if } A \leq x \leq B \\ 0 & \text{otherwise} \end{cases} \quad (4.22)$$

$$E[X] = A + (B - A) \frac{\alpha}{\alpha + \beta}$$

$$\text{Var}[X] = \frac{(B - A)^2 \alpha \beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)}$$

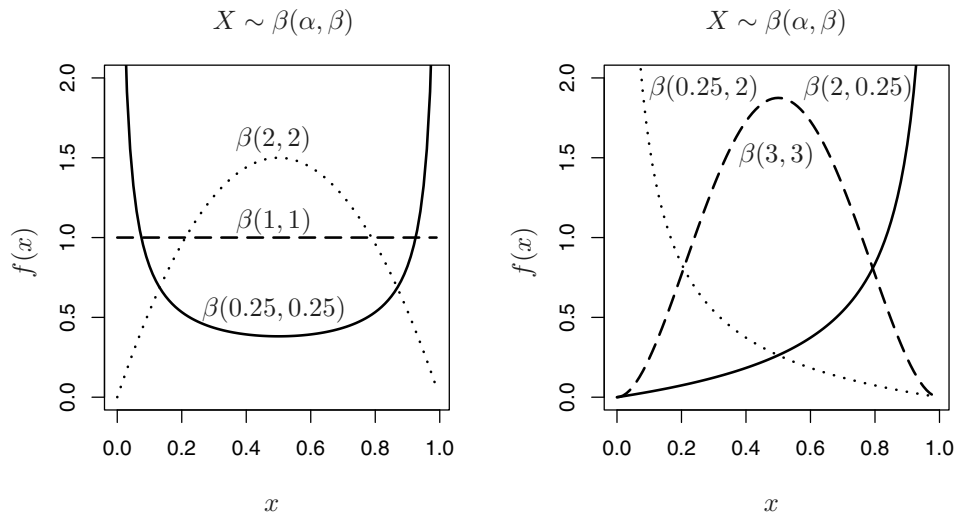


FIGURE 4.11: Illustration of standard $\beta(A = 0, B = 1)$ pdfs for several combinations of α and β

Example 4.23 \triangleright *Beta Distribution: Selling Computers* \triangleleft A wholesale computer distributor has a fixed amount of storage space in his warehouse. The warehouse is restocked with computers every 15 days. The distributor would like more information on the proportion of computers in the warehouse that are sold every 15 days. The warehouse manager claims that the proportion of computers sold can be modeled with a standard beta distribution where $\alpha = 4$ and $\beta = 2$. Compute the expected value for the proportion of computers sold every 15 days. How likely is it that at least 80% of the computers in stock will be sold during a 15 day period?

Solution: Let the random variable X represent the proportion of computers sold in a 15 day period. Since $X \sim \beta(4, 2)$, the expected value from (4.22) yields

$$E[X] = \frac{\alpha}{\alpha + \beta} = \frac{2}{3}.$$

The probability that at least 80% of the computers in the warehouse are sold is

$$\mathbb{P}(X \geq 0.8) = \int_{0.8}^1 \frac{\Gamma(4+2)}{\Gamma(4)\Gamma(2)} x^3(1-x) dx = 20 \int_{0.8}^1 (x^3 - x^4) dx = 0.26272.$$

To compute the last answer with S, use the command `pbeta()` or `integrate()`:

```
> 1 - pbeta(0.8,4,2)
[1] 0.26272

> b42 <- function(x){(gamma(6)/(gamma(4)*gamma(2)))*x^3*(1-x)}
> integrate(b42,0.8,1) # R
0.26272 with absolute error < 2.9e-15

> b42 <- function(x){(gamma(6)/(gamma(4)*gamma(2)))*x^3*(1-x)}
> integrate(b42,0.8,1)$integral # S-PLUS
[1] 0.26272
```

Example 4.24 ▷ *Beta Distribution: Roof My House* ◁ Project managers often use a Program Evaluation and Review Technique (PERT) to manage large scale projects. PERT was actually developed by the consulting firm of Booz, Allen, & Hamilton in conjunction with the United States Navy as a tool for coordinating the activities of several thousands of contractors working on the Polaris missile project. A standard assumption in PERT analysis is that the time to complete any given activity follows a general beta distribution, where A is the optimistic time to complete an activity and B is the pessimistic time to complete the activity. Suppose the time X (in hours) it takes a three man crew to re-roof a single-family house has a beta distribution with $A = 8$, $B = 16$, $\alpha = 2$, and $\beta = 3$. The crew will complete the re-roofing in a single day provided the total time to complete the job is no more than 10 hours. If this crew is contracted to re-roof a single-family house, what is the chance that they will finish the job in the same day?

Solution: To answer the question, find $\mathbb{P}(X \leq 10)$:

$$\begin{aligned} \mathbb{P}(X \leq 10) &= \int_8^{10} \frac{1}{8} \cdot \frac{\Gamma(5)}{\Gamma(2)\Gamma(3)} \left(\frac{x-8}{8}\right) \left(\frac{16-x}{8}\right)^2 dx \\ &= \frac{\Gamma(5)}{8^4\Gamma(2)\Gamma(3)} \int_8^{10} (x-8)(16-x)^2 dx \\ &= \frac{24}{4096 \cdot 1 \cdot 2} \int_8^{10} (512x - 40x^2 + x^3 - 2048) dx \\ &= \frac{3}{1024} \cdot \left(256x^2 - \frac{40}{3}x^3 + \frac{x^4}{4} - 2048x\right) \Bigg|_8^{10} \\ &= \frac{3}{1024} \cdot \frac{268}{3} = 0.2617 \end{aligned}$$

To compute the last answer with S, use the command `integrate()`:

```
> GB <- function(x)
{(1/8)*(gamma(5)/(gamma(2)*gamma(3)))*((x-8)/8)*((16-x)/8)^2}
> integrate(GB, 8, 10) # R
0.2617188 with absolute error < 2.9e-15

> GB <- function(x)
{(1/8)*(gamma(5)/(gamma(2)*gamma(3)))*((x-8)/8)*((16-x)/8)^2}
> integrate(GB, 8, 10)$integral # S-PLUS
[1] 0.2617188
```

To solve the problem with `pbeta()`, enter

```
> A <- 8
> B <- 16
> x <- 10
> pbeta((x-A)/(B-A),2,3)
[1] 0.2617188
```

■

4.3.7 Normal (Gaussian) Distribution

The **normal** or **Gaussian** distribution is more than likely the most important distribution in statistical applications. This is due to the fact that many numerical populations have distributions that can be approximated with the normal distribution. Examples of distributions following an approximate normal distribution include physical characteristics such as the height and weight of a particular species. Further, certain statistics, such as the mean, follow an approximate normal distribution when certain conditions are satisfied. The **pdf**, mean, variance, and **mgf** for a normal random variable X with mean μ and variance σ^2 are provided in (4.23). The **pdf** for a normal random variable has an infinite number of centers and spreads, depending on both μ and σ , respectively. Although there are an infinite number of possible normal distributions, all normal distributions have a bell shape that is symmetric around the distribution's mean. Figure 4.12 on the next page illustrates three normal distributions with identical means, μ , and increasing variances as the distributions are viewed from left to right. The standard deviation in a normal distribution is the horizontal distance from the center of the distribution to the point on the density curve where the curve changes from concave down to concave up (point of inflection). Small values of σ produce distributions that are relatively close to the distribution's mean. On the other hand, values of σ that are large produce distributions that are quite spread out around the distribution's mean.

<p>Normal Distribution</p> <p style="text-align: center;">$X \sim N(\mu, \sigma)$</p> $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty,$ <p style="text-align: center;">where $-\infty < \mu < \infty$ and $0 < \sigma < \infty$.</p> $E[X] = \mu$ $Var[X] = \sigma^2$ $M_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}$	(4.23)
--	--------

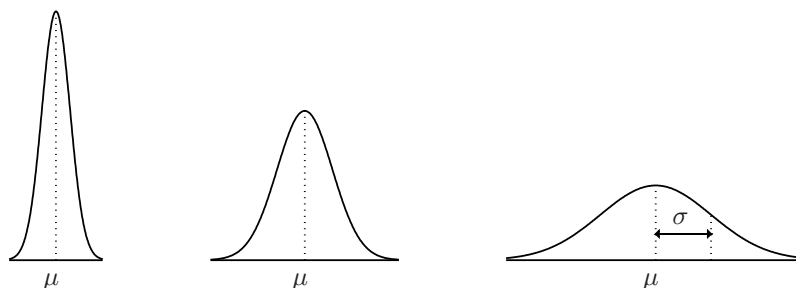


FIGURE 4.12: Three normal distributions, each with an increasing σ value as read from left to right

The **cdf** for a normal random variable, X , with mean, μ , and standard deviation, σ , is

$$F(x) = \mathbb{P}(X \leq x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^x e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt. \quad (4.24)$$

A normal random variable with $\mu = 0$ and $\sigma = 1$, often denoted Z , is called a **standard normal** random variable. The **cdf** for the standard normal distribution, given in (4.26), is computed by first standardizing the random variable X , where $X \sim N(\mu, \sigma)$, using the change of variable formula,

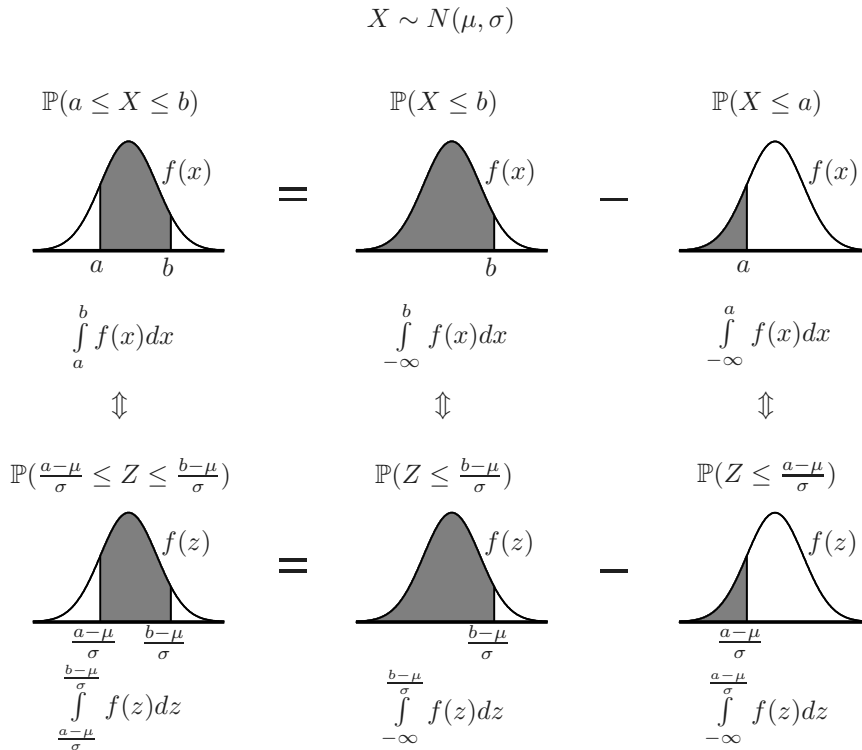
$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1). \quad (4.25)$$

$$F(x) = \mathbb{P}(X \leq x) = \mathbb{P}\left(Z \leq \frac{x - \mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x - \mu}{\sigma}} e^{-\frac{z^2}{2}} dz \quad (4.26)$$

Neither the integral for (4.26) nor the integral for (4.24) can be computed with standard techniques of integration. However, (4.26) has been numerically evaluated and tabled. Further, any normal random variable can be converted to a standard normal random variable using (4.25). The process of computing $\mathbb{P}(a \leq X \leq b)$, where $X \sim N(\mu, \sigma)$, is graphically illustrated in Figure 4.13 on the following page.

Throughout the text, the convention z_α is used to represent the value of the standard normal random variable Z that has α of its area to the left of said value. In other words, $\mathbb{P}(Z < z_\alpha) = \alpha$. Another notation that is also used in the text is $\Phi(z_\alpha) = \alpha$. Basically, the $\Phi(\text{value})$ is the same as $\mathbb{P}(Z < \text{value})$. That is, Φ is the **cdf** of the standard normal distribution. Likewise, $\Phi^{-1}(\alpha) = z_\alpha$. The Φ notation for the **cdf** and inverse **cdf** is used more in Chapter 10.

To find the numerical value of X_α , where $X \sim N(\mu, \sigma)$ and α is the area (or probability) to the left of the value X_α , use the **S** command `qnorm(p, mean=MValue, sd=SValue)`, where **p** is the area or probability (this is equivalent to α) to the left of X_α , **MValue** is the value of the mean, and **SValue** is the value of the standard deviation. Note that if one is dealing with the standard normal distribution, the **mean=MValue** or **sd=SValue** arguments are not needed.

FIGURE 4.13: Graphical representation for computing $\mathbb{P}(a \leq X \leq b)$

Example 4.25 Scores on a particular standardized test follow a normal distribution with a mean of 100 and standard deviation of 10.

- What is the probability that a randomly selected individual will score between 90 and 115?
- What score does one need to be in the top 10%?
- Find the constant c such that $\mathbb{P}(105 \leq X \leq c) = 0.10$.

Solution: Historically, normal distributions had to be standardized and the values of probabilities looked up in tables. Though this is no longer the case, this example shows how to standardize X and to use the `S` command `pnorm()` with a standard normal random variable to “look up” probabilities to the left of given values. Understanding the standard normal, $Z \sim N(0, 1)$, and the probabilities associated with different values from this distribution gives the student an intuition about other normal distributions whose mean and standard deviation are something other than 0 and 1.

(a) To find $\mathbb{P}(90 \leq X \leq 115)$, first draw a picture representing the desired area such as the one in Figure 4.14 on page 156. Note that finding the area between 90 and 115 is equivalent to finding the area to the left of 115 and from that area, subtracting the area to the left of 90. In other words,

$$\mathbb{P}(90 \leq X \leq 115) = \mathbb{P}(X \leq 115) - \mathbb{P}(X \leq 90).$$

To find $\mathbb{P}(X \leq 115)$ and $\mathbb{P}(X \leq 90)$, one can standardize using (4.25). That is,

$$\mathbb{P}(X \leq 115) = \mathbb{P}\left(Z \leq \frac{115 - 100}{10}\right) = \mathbb{P}(Z \leq 1.5),$$

and

$$\mathbb{P}(X \leq 90) = \mathbb{P}\left(Z \leq \frac{90 - 100}{10}\right) = \mathbb{P}(Z \leq -1.0).$$

Using the S commands `pnorm(1.5)` and `pnorm(-1)`, find the areas to the left of 1.5 and -1.0 to be 0.9332 and 0.1586, respectively. Consequently,

$$\begin{aligned}\mathbb{P}(90 \leq X \leq 115) &= \mathbb{P}(-1.0 \leq Z \leq 1.5) \\ &= \mathbb{P}(Z \leq 1.5) - \mathbb{P}(Z \leq -1.0) \\ &= 0.9332 - 0.1587 = 0.7745.\end{aligned}$$

(b) Finding the value c such that 90% of the area is to its left is equivalent to finding the value c such that 10% of its area is to the right. That is, finding the value c that satisfies $\mathbb{P}(X \leq c) = 0.90$ is equivalent to finding the value c such that $\mathbb{P}(X \geq c) = 0.10$. Since the `qnorm()` function refers to areas to the left of a given value by default, solve

$$\mathbb{P}(X \leq c) = \mathbb{P}\left(Z = \frac{X - 100}{10} \leq \frac{c - 100}{10}\right) = 0.90 \text{ for } c.$$

Using `qnorm(.9)`, find the Z value (1.2816) such that 90% of the area in the distribution is to the left of that value. Consequently, to be in the top 10%, one needs to be more than 1.2816 standard deviations above the mean:

$$\begin{aligned}\frac{c - 100}{10} &\stackrel{\text{set}}{=} 1.2816 \\ \text{and solve for } c &\Rightarrow c = 112.816.\end{aligned}$$

To be in the top 10%, one needs to score 112.816 or higher.

(c) $\mathbb{P}(105 \leq X \leq c) = 0.10$ is the same as

$$\mathbb{P}(X \leq c) = 0.10 + \mathbb{P}(X \leq 105) = 0.10 + \mathbb{P}\left(Z \leq \frac{105 - 100}{10}\right).$$

Using `pnorm(.5)`,

$$\mathbb{P}\left(Z \leq \frac{105 - 100}{10}\right) = \mathbb{P}(Z \leq 0.5) = 0.6915.$$

It follows then that $\mathbb{P}(X \leq c) = 0.7915$. Using `qnorm(.7915)` gives 0.8116:

$$\mathbb{P}(X \leq c) = \mathbb{P}\left(Z = \frac{X - 100}{10} \leq \frac{c - 100}{10}\right) = 0.7915$$

$$\text{is found by solving } \frac{c - 100}{10} = 0.8116 \Rightarrow c = 108.116$$

Note that a Z value of 0.8116 has 79.15% of its area to the left of that value.

The following S commands can be used to solve (a),(b), and (c), respectively:

(a) $\mathbb{P}(90 \leq X \leq 115)$

```
> pnorm(115,100,10) - pnorm(90,100,10)
[1] 0.7745375
```

(b) $\mathbb{P}(X \leq c) = 0.90$

```
> qnorm(.90,100,10)
[1] 112.8155
```

(c) $\mathbb{P}(105 \leq X \leq c) = 0.10$

```
> qnorm(.10 + pnorm(105,100,10),100,10)
[1] 108.1151
```

■

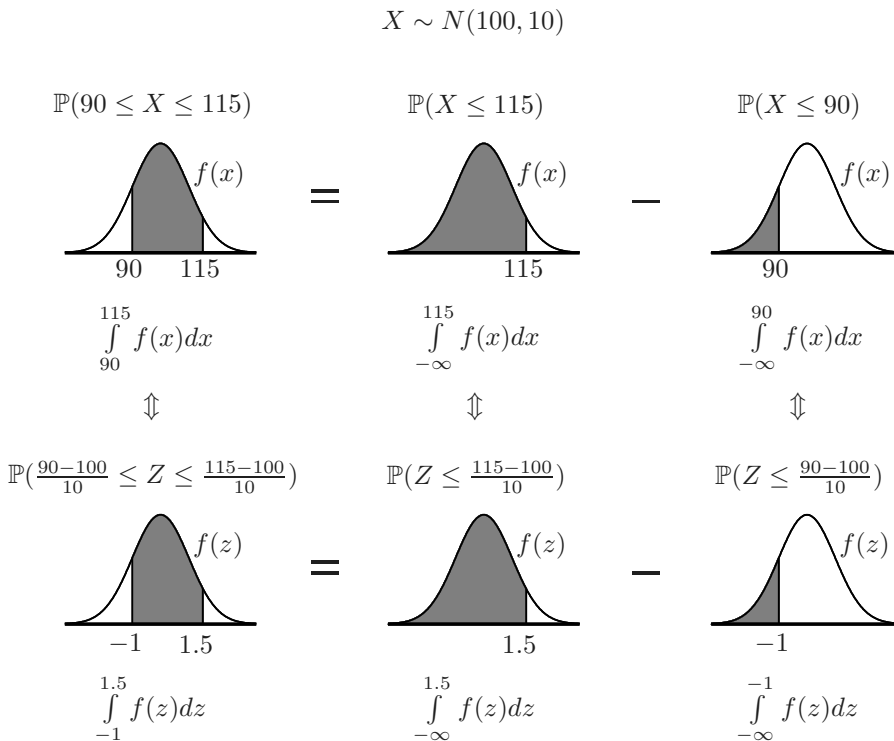


FIGURE 4.14: Graphical representation for finding $\mathbb{P}(90 \leq X \leq 115)$ given $X \sim N(100, 10)$

Example 4.26 ▷ *Normal Distribution: Cell Phone Components* ◁ Most mobile appliances today allow the consumer to switch from the built-in speaker and microphone to an external source. A manufacturer of cell phones wants to package an external speaker and microphone for hands-free operation. A new company has patented a component that

allows the on-resistance flatness for both the microphone and speaker to be lower than ever before. The cell phone company requires that the on-resistance flatness be less than 0.7 ohms (Ω). If it is known that 50% of the components from the new company have an ohm rating of 0.5 Ω or less, 10% have an ohm rating of 0.628 Ω or greater, and the distribution of the ohm ratings is normal, then:

- Find the mean and standard deviation for the distribution of the ohm rating of the components.
- If a component is selected at random, what is the probability that its on-resistance flatness will be less than 0.7 Ω ?
- If 20 components are selected at random, what is the probability that at least 19 components will have on-resistance flatness values less than 0.7 Ω ?

Solution: Let X = the ohm rating of the patented components.

(a) Because a normal distribution is symmetric, the mean equals the median. It is known that 50% of the components have an ohm rating of 0.5 Ω or less, so $\mu_X = 0.5$. To calculate the standard deviation of the components' ohm ratings, use the fact that "10% have an ohm rating of 0.628 Ω or greater."

This means that $\mathbb{P}(X \leq 0.628) = 0.9$,

which implies $\mathbb{P}\left(Z = \frac{X - 0.5}{\sigma} \leq \frac{.628 - .5}{\sigma}\right) = 0.9$.

Because $\mathbb{P}(Z \leq 1.28) = 0.9$, set $\frac{0.628 - 0.5}{\sigma} = 1.28$

and solve for σ . $\frac{0.628 - 0.5}{1.28} = \sigma$

Therefore $\sigma = 0.1$.

(b) Calculate that the probability a component has an on-resistance flatness less than 0.7 Ω :

$$\begin{aligned}\mathbb{P}(X \leq 0.7) &= \mathbb{P}\left(Z = \frac{X - 0.5}{.1} \leq \frac{0.7 - 0.5}{0.1}\right) \\ &= \mathbb{P}(Z \leq 2) \\ &= 0.97725\end{aligned}$$

The answer computed with S is

```
> p <- pnorm(0.7, 0.5, 0.1)
> p
[1] 0.9772499
```

(c) Calculate the probability that at least 19 of the 20 components will have an on-resistance flatness value less than 0.7 Ω . Let $Y \sim \text{Bin}(20, 0.97725)$.

$$\mathbb{P}(Y \geq 19) = \sum_{i=19}^{20} \binom{20}{i} (0.97725)^i (1 - 0.97725)^{20-i} = 0.9250$$

To compute the answer with S, type

```
> sum(dbinom(19:20,20, p))
[1] 0.9249673
```



Quantile-Quantile Plots for Normal Distributions Many of the techniques presented later in the book assume the underlying distribution is normal. One of the more useful graphical procedures for assessing distributions is the quantile-quantile plot. (Recall from Section 2.7.3 that this graph is also called a Q-Q plot.) To help determine whether the underlying distribution is normal, use the S function `qqnorm()`.

To understand the `qqnorm()` function, one needs to have some understanding of S's `quantile()` function. Recall that the cumulative distribution function (`cdf`) is $F(x) = P(X \leq x)$. The `quantile()` function is the inverse of the `cdf`, where this exists; that is, $Q(u) = F^{-1}(u)$. The `qqnorm()` function works by first computing the quantiles of the points $(i - 1/2)/n$ for the standard normal distribution. The ordered sample values are then plotted against the quantiles. When the resulting plot is linear, it indicates the sample values have a normal distribution. To help assess the linearity of the `qqnorm()` plot, it is often quite helpful to plot a straight line through the 25th and 75th percentiles, also referred to as the first and third quartiles, using the S function `qqline()`, which connects the pair of points (First Quartile Standard Normal, First Quartile Data), (Third Quartile Standard Normal, Third Quartile Data).

For example, consider the values stored in the variable `scores` of the data frame `Score` and reported in Table 4.2 which are the scores a random sample of 20 college freshmen received on a standardized test. The points $(i - 1/2)/n$ are calculated as

$$(1 - 1/2)/20 = 0.025, (2 - 1/2)/20 = 0.075, \dots, (20 - 1/2)/20 = 0.975,$$

while the corresponding standard normal quantiles of $\{0.025, 0.075, \dots, 0.975\}$ are computed with `qnorm()` to be $\{-1.96, -1.44, \dots, 1.96\}$, respectively. The S function `qqnorm()` plots the quantiles $\{-1.96, -1.44, \dots, 1.96\}$ versus the ordered values in the sample, $\{87, 90, \dots, 119\}$ as shown in Figure 4.15 on the next page. The pair of points (First Quartile Standard Normal, First Quartile Data), (Third Quartile Standard Normal, Third Quartile Data) is $(-0.637, 96.75)$ and $(0.637, 106.25)$, respectively. Note how the line in Figure 4.15 on the facing page created using the S function `qqline()` goes through the points $(-0.637, 96.75)$ and $(0.637, 106.25)$.

Table 4.2: Standardized scores (data frame `Score`)

119	107	96	107	97	103	94	106	87	112
99	99	90	106	110	99	105	100	100	94

To compute the pairs of values plotted in an S quantile-quantile plot, issue the following commands:

```
> attach(Score)
> par(pty="s")
> X <- (1:20-1/2)/20
> Xs <- qnorm(X)
> Ys <- sort(scores)
> plot(Xs, Ys)
```

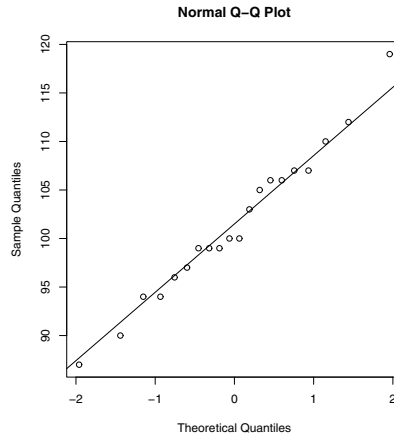


FIGURE 4.15: Quantile-quantile plot of the standardized test scores of 20 randomly selected college freshmen

```
> quantile(Xs, c(0.25, 0.75))
      25%      75%
-0.6371739  0.6371739
> quantile(Ys, c(0.25, 0.75))
      25%      75%
 96.75 106.25
> detach(Score)
```

Generally, the command `qqnorm()` is used to generate the pairs of values that are plotted for a normal quantile-quantile plot, while the command `qqline()` adds a line to a normal quantile-quantile plot that passes through the first and third quartiles. The commands `qqnorm(scores)` and `qqline(scores)` were used to create Figure 4.15.

It is possible to tell from a quantile-quantile plot whether the distribution has shorter or longer tails than a normal distribution. In addition, the quantile-quantile plot will show whether a distribution is skewed and in which direction the distribution is skewed. The right quantile-quantile plots in Figure 4.16 on the following page illustrate how distributions that have a positive skew will appear as upward opening U shapes in the quantile-quantile plot, while distributions with a negative skew have downward facing U shapes. The left quantile-quantile plots in Figure 4.16 on the next page illustrate how distributions that have short tails relative to the normal distribution will have an S shape while distributions with tails longer than the normal distribution will have an inverted S shape.

The graphs in Figure 4.16 can be slightly misleading in the sense that they were constructed from large data sets ($n = 500$). When n is smaller, reading a quantile-quantile plot is slightly more challenging. However, the plotted values still need to fall close to a straight line. One way to train the eye with the quantile-quantile plot is to use simulation to generate data from a normal distribution for various values of n and observe the resulting quantile-quantile plots. When this is done, what one realizes is that for small values of n , even when sampling from a normal distribution, the resulting quantile-quantile plot is not always linear. The function `ntester()`, available in the PASWR package, demonstrates how samples ($n < 5000$) from a normal distribution that have the same sample size as the actual data can appear in quantile-quantile plots. One is strongly encouraged to run this

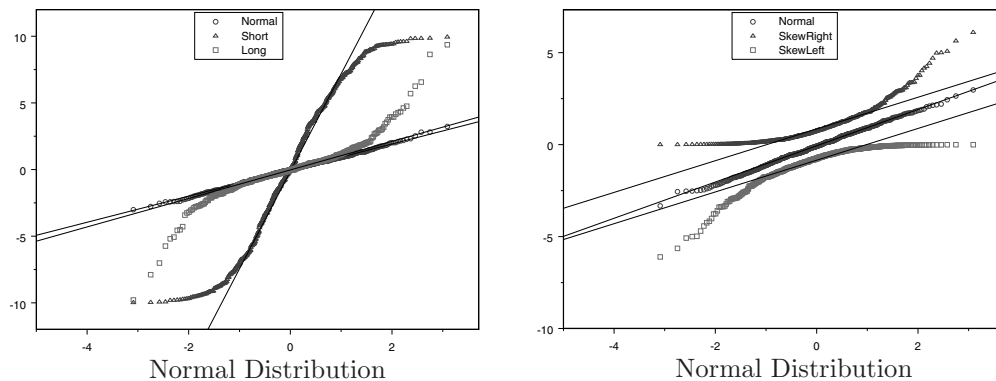


FIGURE 4.16: Superimposed quantile-quantile plots for simulated data from a skew left, skew right, and normal distribution (on the right) and from a short-tailed, long-tailed, and normal distribution (on the left)

function before finalizing the assessment about the normality of a smaller sized sample. The results from using `ntester()` on the standardized test scores from Table 4.2 on page 158 are shown in Figure 4.17 on the next page. Note that the actual data are the center normal quantile-quantile plot and all of the surrounding quantile-quantile plots are for simulated normal data having the same sample size as the center plot. One should pay close attention to how variable the eight surrounding graphs can be even when the data are coming from a normal distribution. If the data are no more variable than the surrounding plots, it should be safe to assume they are normal.

It is often helpful to look at several graphs at once when assessing the general shape of a distribution. The function `EDA()` in the `PASWR` package displays a histogram, a density plot, a boxplot, and a normal quantile-quantile plot of a numeric variable as well as computing various numerical summaries that are returned in the console. In order to allow the user to focus strictly on the resulting shapes, no measurement scales are given in the graphical output. Figure 4.18 on the facing page shows the graphical results from using `EDA(scores)`. All four graphs in Figure 4.18 confirm normality as a reasonable assumption for the distribution of the variable `scores`.

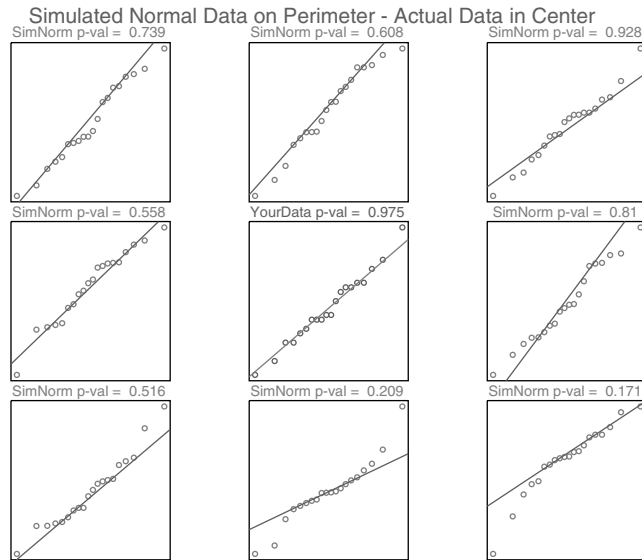


FIGURE 4.17: Resulting quantile-quantile plots using the function `ntester()` on the standardized test scores from Table 4.2 on page 158

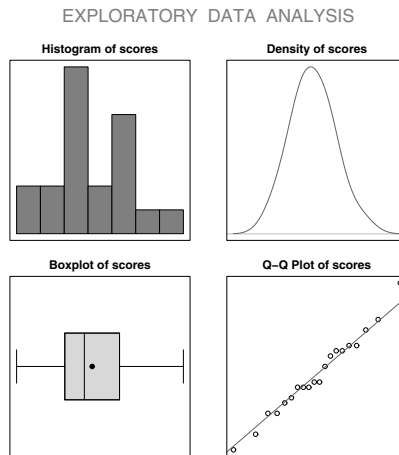


FIGURE 4.18: Graphical results from `EDA(scores)`

4.4 Problems

1. Derive the mean and variance for the discrete uniform distribution.

(Hints: $\sum_{i=1}^n x_i = \frac{n(n+1)}{2}$; $\sum_{i=1}^n x_i^2 = \frac{n(n+1)(2n+1)}{6}$, when $x_i = 1, 2, \dots, n$.)

2. Construct a plot for the probability mass function and the cumulative probability distribution of a binomial random variable $Bin(n = 8, \pi = 0.3)$. Find the smallest value of k such that $\mathbb{P}(X \leq k) \geq 0.44$ when $X \sim Bin(n = 8, \pi = 0.7)$. Calculate $\mathbb{P}(Y \geq 3)$ if $Y \sim Bin(20, 0.2)$.
3. Let X be a Poisson random variable with mean equal to 2. Find $\mathbb{P}(X = 0)$, $\mathbb{P}(X \geq 3)$, and $\mathbb{P}(X \leq k) \geq 0.70$.
4. Let X be an exponential random variable $Exp(\lambda = 3)$. Find $\mathbb{P}(2 < X < 6)$.
5. Fix the seed value at 500, and generate a random sample of size $n = 10000$ from a $Unif(0,1)$ distribution. Calculate the sample mean and the sample variance. Are your answers within 2% of the theoretical values for the mean and variance of a $Unif(0,1)$ distribution?
6. Fix the seed value at 50, and generate a random sample of size $n = 10000$ from an exponential distribution with $\lambda = 2$. Create a density histogram and superimpose the histogram with a theoretical $Exp(\lambda = 2)$ distribution. Calculate the sample mean and the sample variance of the randomly generated values. Are your answers within 2% of the theoretical values for the mean and variance of an $Exp(\lambda = 2)$ distribution?
7. The Laplace distribution, also known as a double exponential, has a **pdf** given by

$$f(x) = \frac{\lambda}{2} \cdot e^{-\lambda|x-\mu|}, \text{ where } -\infty < x < \infty, -\infty < \mu < \infty, \lambda > 0.$$

- (a) Find the theoretical mean and variance of a Laplace distribution. (Hint: Integrals of absolute values should be done as a positive and negative part, in this case, with limits from $-\infty$ to μ and from μ to ∞ .)
- (b) Let X_1 and X_2 be independent exponential random variables, each with parameter λ . The distribution of $Y = X_1 - X_2$ is a Laplace distribution with a mean of zero and a standard deviation of $\sqrt{2}/\lambda$. Set the seed equal to 3, and generate 25,000 X_1 values from an $Exp(\lambda = \frac{1}{2})$ and 25,000 X_2 values from another $Exp(\lambda = \frac{1}{2})$ distribution. Use these values to create the simulated distribution of $Y = X_1 - X_2$.
- (i) Superimpose a Laplace distribution over a density histogram of the Y values. (Hint: The R function `curve()` can be used to superimpose the Laplace distribution over the density histogram.)
- (ii) Is the mean of Y within 0.02 of the theoretical mean?
- (iii) Is the variance of Y within 2% of the theoretical variance?
8. Let X be a normal random variable $N(\mu = 7, \sigma = 3)$. Calculate $\mathbb{P}(X > 7.1)$. Find the value of k such that $\mathbb{P}(X < k) = 0.8$.
9. Let X be a normal random variable $N(\mu = 3, \sigma = \sqrt{0.5})$. Calculate $\mathbb{P}(X > 3.5)$.
10. Let X be a gamma random variable $\Gamma(\alpha = 2, \lambda = 6)$. Find the value a such that $\mathbb{P}(X < a) = 0.95$.

11. If X is the number of 3's that appear when 60 dice are tossed, what is the $E(X^2)$?
12. An importing company knows that 80% of its imported Chinese socks are suitable for sale. If a sample of 60 pairs is drawn at random, find the probability that a percentage between 70% and 90% (inclusive) of the sample is suitable for sale.
13. It is known that 3% of the seeds of a certain variety of tomato do not germinate. The seeds are sold in individual boxes that contain 20 seeds per box with the guarantee that at least 18 seeds will germinate. Find the probability that a randomly selected box does not fulfill the aforementioned requirement.
14. Traffic volume is an important factor for determining the most cost effective method to surface a road. Suppose that the average number of vehicles passing a certain point on a road is 2 every 30 seconds.
 - (a) Find the probability that more than 3 cars will pass the point in 30 seconds.
 - (b) What is the probability that more than 10 cars pass the point in 3 minutes?
15. The retaining wall of a dam will break if it is subjected to the pressure of two floods. If the average number of floods in a century is two, find the probability that the retaining wall lasts more than 20 years.
16. A particular competition shooter hits his targets 70% of the time with any pistol. To prepare for shooting competitions, this individual practices with a pistol that holds 5 bullets on Tuesday, Thursday, and Saturday, and a pistol that holds 7 bullets the other days. If he fires at targets until the pistol is empty, find the probability that he hits only one target out of the bullets shot in the first round of bullets in the pistol he is carrying that day. In this case, what is the probability that he used the pistol with 7 bullets?
17. A binomial, $Bin(n, \pi)$, distribution can be approximated by a normal distribution, $N(n\pi, \sqrt{n\pi(1-\pi)})$, when $n\pi > 10$ and $n(1-\pi) > 10$. The Poisson distribution can also be approximated by a normal distribution $N(\lambda, \sqrt{\lambda})$ if $\lambda > 10$. Consider a sequence from 7 to 25 of a variable X (binomial or Poisson) and show that for $n = 80$, $\pi = 0.2$, and $\lambda = 16$ the aforementioned approximations are appropriate. The normal approximation to a discrete distribution can be improved by adding 0.5 to the normal random variable when finding the area to the left of said random variable. Specifically, create a table showing $\mathbb{P}(X \leq x)$ for the range of X for the three distributions and a graph showing the density of the normal distribution with vertical lines at $X - .1$ and $X + .1$ showing $\mathbb{P}(X = x)$ for the binomial and Poisson distributions, respectively.
18. Verify that if k/N is small (≤ 0.1) and $N = m+n$ is large, a hypergeometric distribution, $Hyper(m, n, k)$, can be adequately approximated by a $Bin(n = k, \pi = m/N)$ distribution. Compute the probabilities for each distribution using the values $n = 20$, $m = 300$, $k = 10$. Show the numerical results to three decimal places as well as a graph depicting the probabilities of the hypergeometric distribution with a vertical line and the probabilities of the binomial distribution in the same plot with an open circle.
19. In 1935, Fisher described the following experiment in his book, *Design of Experiments*: A friend of Fisher's said that when she drank tea with milk, she was able to determine if the tea was poured first or if the milk was poured first. Find the probability that Fisher's colleague guesses 3 cups in which milk has been added before tea, given that in 4 out of 8 cups, milk has been added before tea.

20. Consider the function $g(x) = (x - a)^2$, where a is a constant and $E[(X - a)^2]$ is finite. Find a so that $E[(X - a)^2]$ is minimized.
21. Suppose the percentage of drinks sold from a vending machine are 80% and 20% for soft drinks and bottled water, respectively.
- What is the probability that on a randomly selected day, the first soft drink is the fourth drink sold?
 - Find the probability that exactly 1 out of 10 drinks sold is a soft drink.
 - Find the probability that the fifth soft drink is the seventh drink sold.
 - Verify empirically that $\mathbb{P}(\text{Bin}(n, \pi) \leq r - 1) = 1 - \mathbb{P}(\text{NB}(r, \pi) \leq (n - r))$, with $n = 10$, $\pi = 0.8$, and $r = 4$.
22. The hardness of a particular type of sheet metal sold by a local manufacturer has a normal distribution with a mean of 60 micra and a standard deviation of 2 micra.
- This type of sheet metal is said to conform to specification provided its hardness measure is between 57 and 65 micra. What percent of the manufacturer's sheet metal can be expected to fall within the specification?
 - A building contractor agrees to purchase metal from the local metal manufacturer at a premium price provided four out of four randomly selected pieces of metal test between 57 and 65 micra. What is the probability the building contractor will purchase metal from the local manufacturer and pay a premium price?
 - If an acceptable sheet of metal is one whose hardness is not more than c units away from the mean, find c such that 97% of the sheets are acceptable.
 - Find the probability that at least 10 out of 20 sheets have a hardness greater than 60.
23. The weekly production of a banana plantation can be modeled with a normal random variable that has a mean of 5 tons and a standard deviation of 2 tons.
- Calculate the mean number of weeks in which the production is greater than the third quartile.
 - Find the probability that, in at most 1 out of the 8 randomly chosen weeks, the production has been less than 3 tons.
 - Find the probability that at least 3 weeks are needed to obtain a production greater than 10 tons.
24. The lifetime of a certain engine follows a normal distribution with mean and standard deviation of 10 and 3.5 years, respectively. The manufacturer replaces all catastrophic engine failures within the guarantee period free of charge. If the manufacturer is willing to replace no more than 4% of the defective engines, what is the largest guarantee period the manufacturer should advertise?
25. A bank has 50 deposit accounts with €25,000 each. The probability of having to close a deposit account and then refund the money in a given day is 0.01. If account closings are independent events, how much money must the bank have available to guarantee it can refund all closed accounts in a given day with probability greater than 0.95?

26. The vendor in charge of servicing coffee dispensers is adjusting the one located in the department of statistics. To maximize profit, adjustments are made so that the average quantity of liquid dispensed per serving is 200 milliliters per cup. Suppose the amount of liquid per cup follows a normal distribution and 5.5% of the cups contain more than 224 milliliters.
- Find the probability that a given cup contains between 176 and 224 milliliters.
 - If the machine can hold 20 liters of liquid, find the probability that the machine must be replenished before dispensing 99 cups.
 - If 6 random samples of 5 cups are drawn, what is the probability that the sample mean is greater than 210 milliliters in at least 2 of them?
27. The mean number of calls a tow truck company receives during a day is 5 per hour. Find the probability that a tow truck is requested more than 4 times per hour in a given hour. What is the probability the company waits for less than 1 hour before the tow truck is requested 3 times?
28. The pill weight for a particular type of vitamin follows a normal distribution with a mean of 0.6 grams and a standard deviation of 0.015 grams. It is known that a particular therapy consisting of a box of vitamins with 125 pills is not effective if more than 20% of the pills are under 0.58 grams.
- Find the probability that the therapy with a box of vitamins is not effective.
 - A supplement manufacturer sells vitamin bottles containing 125 vitamins per bottle with 50 bottles per box with a guarantee that at least 47 bottles per box weigh more than 74.7 grams. Find the probability that a randomly chosen box does not meet the guaranteed weight.
29. A canning industry uses tins with weight equal to 20 grams. The tin is placed on a scale and filled with red peppers until the scale shows the weight μ . Then, the tin contains Y grams of peppers. If the scale is subject to a random error $X \sim N(0, \sigma = 10)$,
- How is Y related to X and μ ?
 - What is the probability distribution of the random variable Y ?
 - Calculate the value μ such that 98% of the tins contain at least 400 grams of peppers.
 - Repeat the exercise assuming the weight of the tins to be a normal random variable $W \sim N(20, \sigma = 5)$.
30. In the printing section of a plastics company, a machine receives on average 6 buckets per minute to be painted. The machine has been out of service for 90 seconds due to a power failure.
- Find the probability that more than 8 buckets remain unpainted.
 - Find the probability that the first bucket, after the electricity is restored, arrives before 10 seconds have passed.
31. Give a general expression to calculate the quantiles of a Weibull random variable.
32. A used-car salesman offers a guarantee period of one year for his cars. He knows that the distribution of the elapsed time until the first breakdown occurs follows a Weibull distribution, $Weib(3, 25)$. If the salesman expects to sell 50 cars per year, and the repair cost per car is on average 800 dollars, what is the mean cost of the guarantee?

33. Let X be a random variable with probability density function

$$f(x) = 3 \left(\frac{1}{x} \right)^4, \quad x \geq 1.$$

- Fix the seed at 98 (`set.seed(98)`), and generate a random sample of size $n = 10000$ from X 's distribution. Compute the mean, variance, and coefficient of skewness for the random sample.
- Obtain the theoretical mean, variance, and coefficient of skewness of X .
- How close are the estimates in (a) to the theoretical values in (b)?

34. Let X be a random variable with probability density function

$$f(x) = \theta \left(\frac{1}{x} \right)^{\theta+1}, \quad x \geq 1, \theta > 1.$$

- Verify that the area under $f(x)$ is 1.
- Fix the seed at 42 (`set.seed(42)`), and generate 10000 realizations of X with $\theta = 2$. What are the mean and variance of the random sample?
- Calculate the theoretical mean and variance of X .
- How close are the estimates in (b) to the theoretical values in (c)?
- Find the cumulative density function.
- What is $\mathbb{P}(X \leq 3)$?

35. Let X be a random variable with probability density function

$$f(x) = \frac{4}{3}x(2 - x^2), \quad 0 \leq x \leq 1.$$

- Verify that the area under $f(x)$ is 1.
- Fix the seed at 13 (`set.seed(13)`), and generate 10000 realizations of X . What are the mean and variance of the random sample?
- Calculate the theoretical mean and variance of X .
- How close are the estimates in (b) to the theoretical values in (c)?
- Find the cumulative density function.
- What is $\mathbb{P}(X > .75)$?

36. Let X be a random variable with probability density function

$$f(x) = (\theta + 1)(1 - x)^\theta, \quad 0 \leq x \leq 1, \theta > 0.$$

- Verify that the area under $f(x)$ is 1.
- Fix the seed at 80 (`set.seed(80)`), and generate 10000 realizations of X with $\theta = 2$. What are the mean and variance of the random sample?
- Calculate the theoretical mean and variance of X .
- How close are the estimates in (b) to the theoretical values in (c)?
- Find the cumulative density function.

(f) What is $\mathbb{P}(X \leq .25)$?

37. Let X be a random variable with probability density function

$$f(x) = 3\pi\theta x^2 e^{-\theta\pi x^3}, \quad x \geq 0.$$

- Verify that the area under $f(x)$ is 1.
- Fix the seed at 201 (`set.seed(201)`), and generate 10000 realizations of X with $\theta = 5$. What are the mean and variance of the random sample?
- Calculate the theoretical mean and variance of X .
- How close are the estimates in (b) to the theoretical values in (c)?
- Find the cumulative density function.
- What is $\mathbb{P}(X > 1)$?

38. A copper wire manufacturer produces conductor cables. These cables are of practical use if their resistance lies between 0.10 and 0.13 ohms per meter. The resistance of the cables follows a normal distribution, where 50% of the cables have resistance under 0.11 ohms and 10% have resistance over 0.13 ohms.

- Determine the mean and the standard deviation for cable resistance.
- Find the probability that a randomly chosen cable can be used.
- Find the probability that at least 3 out of 5 randomly chosen cables can be used.

39. Consider the random variable $X \sim Weib(\alpha, \beta)$.

- Find the **cdf** for X .
- Use (4.18) and verify that for $X \sim Weib(\alpha, \beta)$, the hazard function is given by

$$h(t) = \frac{\alpha t^{\alpha-1}}{\beta^\alpha}.$$

40. If $X \sim Bin(n, \pi)$, derive the moment generating function of X and use it to derive the mean and variance of X . The binomial **pdf** can be found on page 117.

41. If $X \sim Bin(n, \pi)$, use the binomial expansion to find the mean and variance of X . To find the variance, use the second factorial moment $E[X(X-1)]$ and note that $\frac{x}{x!} = \frac{1}{(x-1)!}$ when $x > 1$.

42. The speed of a randomly chosen gas molecule in a certain volume of gas is a random variable, V , with probability density function

$$f(v) = \sqrt{\frac{2}{\pi}} \left(\frac{M}{RT} \right)^{\frac{3}{2}} v^2 e^{-\frac{Mv^2}{2RT}} \quad \text{for } v \geq 0$$

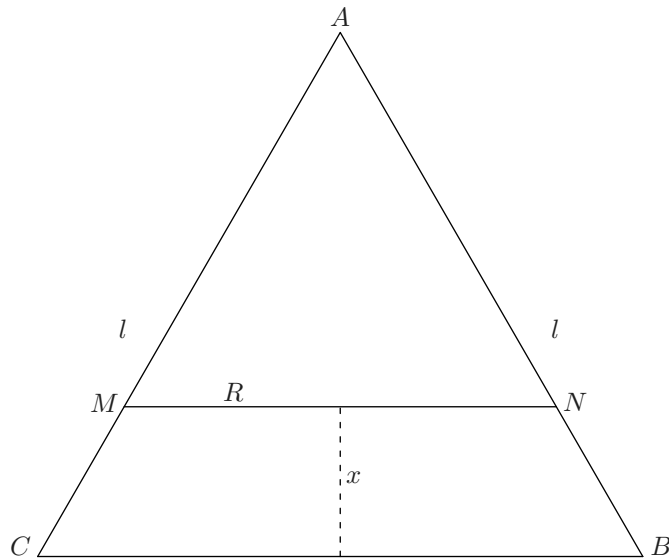
where R is the gas constant ($= 8.3145 \text{ J/mol} \cdot \text{K}$), M is the molecular weight of the gas, and T is the absolute temperature measured in degrees Kelvin.

(Hints:

$$\int_0^\infty x^k e^{-x^2} dx = \frac{1}{2} \Gamma\left(\frac{k+1}{2}\right) \quad \Gamma(\alpha+1) = \alpha \Gamma(\alpha) \quad \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

- Derive a general expression for the average speed of a gas molecule.

- (b) If $1 \text{ J} = 1 \text{ kg} \cdot \text{m}^2 / \text{s}^2$, what are the units for the answer in part (a)?
- (c) Kinetic energy for a molecule is $E_k = \frac{Mv^2}{2}$. Derive a general expression for the average kinetic energy of a molecule.
- (d) The weight of hydrogen is 1.008 g/mol . Note that there are 6.0221415×10^{23} molecules in 1 mole. Find the average speed of a hydrogen molecule at 300°K using the result from part (a).
- (e) Use numerical integration to verify the result from part (d).
- (f) Show the probability density functions for the speeds of hydrogen, helium, and oxygen on a single graph. The molecular weights for these elements are 1.008 g/mol , 4.003 g/mol , and 16.00 g/mol , respectively.
43. Consider the equilateral triangle ABC with side l . Given a randomly chosen point R in the triangle, calculate the cumulative and the probability density functions for the distance from R to the side BC . Construct a graph of the cumulative density function for different values of l . (Hint: The equation of the line CA is $y = \sqrt{3}x$.)



44. In Pamplona, Spain, a tombola organizes different raffles during the festivals. In each raffle, only 2 tickets out of n win a prize. The tickets are sold consecutively, and the

prize is immediately announced when one person wins. Two friends have decided to take part in one of the raffles in the following way: One of them buys the first ticket on sale, and the other one buys the first ticket after the first prize has been announced. Derive the probability that each of them wins a prize. If there are m raffles during the night in which the two friends participate, what is the probability that each of them wins more than one prize?

45. Example 4.4 on page 122 introduced the World Cup Soccer data stored in the data frame `Soccer`. The observed and expected number of goals for a 90 minute game were computed. To verify that the Poisson rate λ is constant, compute the observed and expected number of goals with the time intervals 45, 15, 10, 5, and 1 minute(s). Compute the means and variances for both the observed and expected counts in each time interval. Based on the results, is criterion (3) of the Poisson process on page 120 satisfied? (Note: See the code at the end of the Chapter 4 script for ideas on how to do this.)

Chapter 5

Multivariate Probability Distributions

5.1 Joint Distribution of Two Random Variables

In Sections 3.4.1 and 3.4.5, respectively, both discrete and continuous random variables were defined. However, it stands to reason that many random variables might be defined over the same sample space. In random variable example 1 on page 88, the random variable X was defined as the sum of the numbers from two dice. However, one might also wish to consider “the product of the numbers rolled with the two dice” or “the absolute value of the difference between the numbers rolled with the two dice” as additional random variables that are defined on the same sample space. Another example might be the verbal (X) and quantitative (Y) scores for incoming freshmen at a private college. In this section, a brief overview for both discrete and continuous **pdfs** and **cdfs** of jointly distributed random variables is provided as well as some important properties associated with jointly distributed random variables.

5.1.1 Joint pdf for Two Discrete Random Variables

If X and Y are discrete random variables, the function given by

$$p_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y) \quad (5.1)$$

for each pair of values (x, y) within the domain of X and Y is called the **joint pdf** of X and Y . Any function $p_{X,Y}(x, y)$ can be used as a **joint pdf** provided the following properties are satisfied:

- (i) $p_{X,Y}(x, y) \geq 0$ for all x and y .
- (ii) $\sum_x \sum_y p_{X,Y}(x, y) = 1$.
- (iii) $\mathbb{P}[(X, Y) \in A] = \sum_{(x,y) \in A} p_{X,Y}(x, y)$.

Property (iii) states that when A is composed of pairs of (x, y) values, the probability $\mathbb{P}[(X, Y) \in A]$ is obtained by summing the **joint pdf** over pairs in A .

Example 5.1 ▷ *Joint Distribution: Mathematics Grades* ◁ To graduate with a bachelor of science (B.S.) degree in mathematics, all majors must pass Calculus III and Linear Algebra with a grade of B or better. The population of B.S. graduates in mathematics earned grades as given in Table 5.1 on the next page.

- (a) What is the probability of getting a B or better in Linear Algebra?
- (b) What is the probability of getting a B or better in Calculus III?
- (c) What is the probability of getting a B or better in both Calculus III and Linear Algebra?

Table 5.1: B.S. graduate grades in Linear Algebra and Calculus III

		Linear Algebra		
		A	B	C
Calculus III	A	2	13	6
	B	5	85	40
	C	7	33	9

Solution: The answers are as follows:

(a) Let the random variables X and Y represent the grades in Calculus III and Linear Algebra, respectively. If A represents the pairs of Calculus III and Linear Algebra values such that the grade in Linear Algebra is a B or better, then the probability of getting a B or better in Linear Algebra is written

$$\mathbb{P}[(X, Y) \in A] = \sum_{(x,y) \in A} p_{X,Y}(x, y) = \frac{2 + 5 + 7 + 13 + 85 + 33}{200} = \frac{145}{200}.$$

(b) Let the random variables X and Y represent the grades in Calculus III and Linear Algebra, respectively. If A represents the pairs of Calculus III and Linear Algebra values such that the grade in Calculus III is a B or better, then the probability of getting a B or better in Calculus III is written

$$\mathbb{P}[(X, Y) \in A] = \sum_{(x,y) \in A} p_{X,Y}(x, y) = \frac{2 + 13 + 6 + 5 + 85 + 40}{200} = \frac{151}{200}.$$

(c) Let the random variables X and Y represent the grades in Calculus III and Linear Algebra, respectively. If A represents the pairs of Calculus III and Linear Algebra values such that the grade in both Calculus III and Linear Algebra is a B or better, then the probability of getting a B or better in both Calculus III and Linear Algebra is written

$$\mathbb{P}[(X, Y) \in A] = \sum_{(x,y) \in A} p_{X,Y}(x, y) = \frac{2 + 5 + 13 + 85}{200} = \frac{105}{200}. \quad \blacksquare$$

For any random variables X and Y , the joint **cdf** is defined in (5.2), while the marginal **pdfs** of X and Y , denoted $p_X(x)$ and $p_Y(y)$, respectively, are defined in Equations (5.3) and (5.4):

$$F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y), \quad -\infty < x < \infty, \quad -\infty < y < \infty \quad (5.2)$$

$$p_X(x) = \sum_y p_{X,Y}(x, y) \quad (5.3)$$

$$p_Y(y) = \sum_x p_{X,Y}(x, y) \quad (5.4)$$

In (a) of Example 5.1 on the preceding page, the problem requests the probability of getting a B or better in Linear Algebra. Another way to compute the answer is by adding the two marginals $p_Y(A) + p_Y(B) = \frac{14}{200} + \frac{131}{200} = \frac{145}{200}$. Likewise, (b) of Example 5.1 on the previous page can also be solved with the marginal distribution for X : $p_X(A) + p_X(B) = \frac{21}{200} + \frac{130}{200} = \frac{151}{200}$.

5.1.2 Joint pdf for Two Continuous Random Variables

In Section 3.4.5 on page 93, property (3) for continuous **pdfs** states that the probability the observed value for the random variable X falls in the interval (a, b) is the integral of the **pdf** $f(x)$ over the interval (a, b) . In a similar fashion, the probability that the pair of random variables (X, Y) falls in a two-dimensional region (say A) is obtained by integrating the joint **pdf** over the region A . The joint **pdf** of two continuous random variables is any integrable function $f_{X,Y}(x, y)$ with the following properties:

- (1) $f_{X,Y}(x, y) \geq 0$ for all x and y .
- (2) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx dy = 1$.
- (3) $\mathbb{P}[(X, Y) \in A] = \iint_{(x,y) \in A} f_{X,Y}(x, y) dx dy$.

Property (3) implies that $\mathbb{P}[(X, Y) \in A]$ is the volume of a solid over the region A bounded by the surface $f_{X,Y}(x, y)$.

For any random variables X and Y , the joint **cdf** is defined in (5.5), while the marginal **pdfs** of X and Y , denoted $f_X(x)$ and $f_Y(y)$, respectively, are defined in Equations (5.6) and (5.7):

$$F_{X,Y}(x, y) = \int_{-\infty}^x \int_{-\infty}^y f_{X,Y}(r, s) ds dr, \quad -\infty < x < \infty, \quad -\infty < y < \infty \quad (5.5)$$

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy, \quad -\infty < x < \infty \quad (5.6)$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx, \quad -\infty < y < \infty \quad (5.7)$$

Example 5.2 Given the joint continuous **pdf**

$$f_{X,Y}(x, y) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1, \quad 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

- (a) Find $F_{X,Y}(x = 0.6, y = 0.8)$.
- (b) Find $\mathbb{P}(0.25 \leq X \leq 0.75, 0.1 \leq Y \leq 0.9)$.
- (c) Find $f_X(x)$.

Solution: The answers are as follows:

(a)

$$F_{X,Y}(x = 0.6, y = 0.8) = \int_0^{0.6} \int_0^{0.8} f_{X,Y}(r, s) ds dr = \int_0^{0.6} \int_0^{0.8} 1 ds dr = \int_0^{0.6} 0.8 dr = 0.48$$

(b)

$$\begin{aligned} \mathbb{P}(0.25 \leq x \leq 0.75, 0.1 \leq y \leq 0.9) \\ = \int_{0.25}^{0.75} \int_{0.1}^{0.9} f_{X,Y}(r, s) ds dr = \int_{0.25}^{0.75} \int_{0.1}^{0.9} 1 ds dr = \int_{0.25}^{0.75} 0.8 dr = 0.40 \end{aligned}$$

(c)

$$f_X(x) = \int_0^1 f_{X,Y}(x,y) dy = 1, \quad 0 \leq x \leq 1 \quad \blacksquare$$

Example 5.3 ▷ **Joint PDF** ◁ Find the value c to make $f_{X,Y}(x,y) = cx$ a valid joint pdf for $x > 0$, $y > 0$, and $2 < x + y < 3$.

Solution: The domain of interest is lightly shaded in Figure 5.1. To solve the problem, first compute the volume bounded by $x = 0$, $y = 0$, and $y = 3 - x$ beneath the surface $f_{X,Y}(x,y) = cx$, which is denoted $V1$. Next, find the volume bounded by $x = 0$, $y = 0$, and $y = 2 - x$ beneath the surface $f_{X,Y}(x,y) = cx$, which is denoted $V2$. For $f_{X,Y}(x,y)$ to be a valid pdf, c must be found such that the difference between $V1$ and $V2$ is one.

$$V1 = \int_0^3 \int_0^{3-x} cx dy dx = c \int_0^3 (3x - x^2) dx = c \left[\frac{3x^2}{2} - \frac{x^3}{3} \right]_0^3 = \frac{27c}{6}$$

$$V2 = \int_0^2 \int_0^{2-x} cx dy dx = c \int_0^2 (2x - x^2) dx = c \left[x^2 - \frac{x^3}{3} \right]_0^2 = \frac{8c}{6}$$

$$V1 - V2 = \frac{27c}{6} - \frac{8c}{6} \stackrel{\text{set}}{=} 1 \Rightarrow c = \frac{6}{19}$$

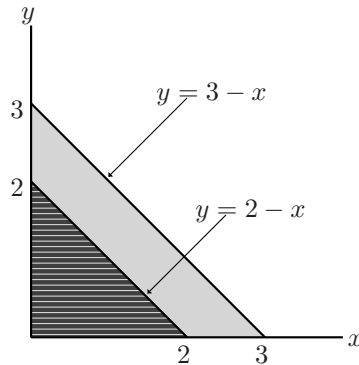


FIGURE 5.1: Graphical representation of the domain of interest for Example 5.3 ▀

5.2 Independent Random Variables

In Section 3.3.6 on page 86, it was shown that two events, E and F , are independent if $\mathbb{P}(E \cap F) = \mathbb{P}(E) \cdot \mathbb{P}(F)$. In a similar fashion, two random variables are independent if for every pair of x and y values, $p_{X,Y}(x,y) = p_X(x) \cdot p_Y(y)$, when X and Y are discrete, or $f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$ when X and Y are continuous.

Example 5.4 Use Table 5.1 on page 172 to decide if the random variables X , grade in Calculus III, and Y , grade in Linear Algebra, are dependent.

Solution: The random variables X and Y are dependent if $p_{X,Y}(x, y) \neq p_X(x) \cdot p_Y(y)$ for any (x, y) . Consider the pair $(x, y) = (A, A)$, that is, an A in both Calculus III and in Linear Algebra.

$$\begin{aligned} p_{X,Y}(A, A) &\stackrel{?}{=} p_X(A) \cdot p_Y(A) \\ \frac{2}{200} &\stackrel{?}{=} \frac{21}{200} \cdot \frac{14}{200} \\ \frac{2}{200} &\neq \frac{21 \times 14}{40,000} \\ 0.01 &\neq 0.00735 \end{aligned}$$

Since $0.01 \neq 0.00735$, the random variables X and Y , the grades in Calculus III and Linear Algebra, respectively, are dependent. It is important to note that the definition of independence requires all the joint probabilities to be equal to the product of the corresponding row and column marginal probabilities. Consequently, if the joint probability of a single entry is not equal to the product of the corresponding row and column marginal probabilities, the random variables in question are said to be dependent. ■

Example 5.5 Are the random variables X and Y in Example 5.2 on page 173 independent? Recall that the **pdf** for Example 5.2 was defined as

$$f_{X,Y}(x, y) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1, \quad 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Solution: Since the marginal **pdf** for X , $f_X(x) = 1$, and the marginal **pdf** for Y , $f_Y(y) = 1$, it follows that X and Y are independent since $f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$ for all x and y . ■

5.3 Several Random Variables

This section examines the joint **pdf** of several random variables by extending the material presented for the joint **pdf** of two discrete random variables and two continuous random variables covered in Section 5.1.1. The joint **pdf** of X_1, X_2, \dots, X_n discrete random variables is any function $p_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) = \mathbb{P}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$ provided the following properties are satisfied:

- $p_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) \geq 0$ for all x_1, x_2, \dots, x_n .
- $\sum_{x_1} \sum_{x_2} \cdots \sum_{x_n} p_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) = 1$.
- $\mathbb{P}[(X_1, X_2, \dots, X_n) \in A] = \sum_{(x_1, x_2, \dots, x_n) \in A} p_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n)$.

The joint **pdf** of X_1, X_2, \dots, X_n continuous random variables is any integrable function $f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n)$ such that following properties are satisfied:

- (a) $f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) \geq 0$ for all x_1, x_2, \dots, x_n .
- (b) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) dx_1 dx_2 \cdots dx_n = 1$
- (c)

$$\mathbb{P}[(X_1, X_2, \dots, X_n) \in A] = \iint_{(x_1, x_2, \dots, x_n) \in A} \cdots \int f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) dx_1 dx_2 \cdots dx_n$$

Independence for several random variables is simply a generalization of the notion for the independence between two random variables. X_1, X_2, \dots, X_n are independent if, for every subset of the random variables, the joint **pdf** of the subset is equal to the product of the marginal **pdfs**. Further, if X_1, X_2, \dots, X_n are independent random variables with respective moment generating functions $M_{X_1}(t), M_{X_2}(t), \dots, M_{X_n}(t)$, then the moment generating function of $Y = \sum_{i=1}^n c_i X_i$ is

$$M_Y(t) = M_{X_1}(c_1 t) \times M_{X_2}(c_2 t) \times \cdots \times M_{X_n}(c_n t). \quad (5.8)$$

In the case where X_1, X_2, \dots, X_n are independent normal random variables, a theorem for the distribution of $Y = a_1 X_1 + \cdots + a_n X_n$, where a_1, a_2, \dots, a_n are constants, is stated.

Theorem 5.1 If X_1, X_2, \dots, X_n are independent normal random variables, with means μ_i and standard deviations σ_i for $i = 1, 2, \dots, n$, the distribution of $Y = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n$, where a_1, a_2, \dots, a_n are constants, is normal with mean $E[Y] = a_1 \mu_1 + a_2 \mu_2 + \cdots + a_n \mu_n$ and variance $\text{Var}[Y] = a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + \cdots + a_n^2 \sigma_n^2$. In other words,

$$Y \sim N \left(\sum_{i=1}^n a_i \mu_i, \sqrt{\sum_{i=1}^n a_i^2 \sigma_i^2} \right)$$

Proof: Since $X_i \sim N(\mu_i, \sigma_i)$, the **mgf** for X_i is $M_{X_i}(t) = e^{\mu_i t + \frac{\sigma_i^2 t^2}{2}}$ using the **mgf** from (4.23). Further, since the X_1, X_2, \dots, X_n are independent,

$$\begin{aligned} M_Y(t) &= M_{X_1}(ta_1) \times M_{X_2}(ta_2) \times \cdots \times M_{X_n}(ta_n) \\ &= e^{t \sum_{i=1}^n a_i \mu_i + t^2 \sum_{i=1}^n \frac{a_i^2 \sigma_i^2}{2}}, \end{aligned}$$

which is the moment generating function for a normal random variable with mean $\sum_{i=1}^n a_i \mu_i$

and variance $\sum_{i=1}^n a_i^2 \sigma_i^2$.

Example 5.6 Use moment generating functions to show that the sum of two independent Poisson random variables is a Poisson random variable.

Solution: First recall that the **mgf** of a Poisson random variable is $M_X(t) = e^{\lambda(e^t - 1)}$. If X is a Poisson random variable with mean λ and Y is a Poisson random variable with mean μ , then $Z = X + Y$ is also a Poisson random variable with mean $\lambda + \mu$ since

$$M_Z(t) = M_X(t) \times M_Y(t) = e^{\lambda(e^t - 1)} \times e^{\mu(e^t - 1)} = e^{(\lambda + \mu)(e^t - 1)}. \quad \blacksquare$$

5.4 Conditional Distributions

Suppose X and Y represent the respective lifetimes (in years) for the male and the female in married couples. If $X = 72$, what is the probability that $Y \geq 75$? In other words, if the male partner of a marriage dies at age 72, how likely is it that the surviving female will live to an age of 75 or more? Questions of this type are answered with conditional distributions. Given two discrete random variables, X and Y , define the conditional **pdf** of X given that $Y = y$ provided that $p_Y(y) > 0$ as

$$p_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}. \quad (5.9)$$

If the random variables are continuous, the conditional **pdf** of X given that $Y = y$ provided that $f_Y(y) > 0$ is defined as

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}. \quad (5.10)$$

In addition, if X and Y are jointly continuous over an interval A ,

$$\mathbb{P}(X \in A|Y = y) = \int_A f_{X|Y}(x|y) dx.$$

Example 5.7 Let the random variables X and Y have a joint **pdf**:

$$f_{X,Y}(x, y) = \begin{cases} \frac{12}{5}x(2 - x - y) & \text{for } 0 < x < 1, 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

Find the **pdf** of X given $Y = y$, for $0 < y < 1$.

Solution: Using the definition for the conditional **pdf** of X given $Y = y$ from (5.10), write

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f_{X,Y}(x, y)}{f_Y(y)} = \frac{f_{X,Y}(x, y)}{\int_{-\infty}^{\infty} f_{X,Y}(x, y) dx} = \frac{x(2 - x - y)}{\int_0^1 x(2 - x - y) dx} \\ &= \frac{x(2 - x - y)}{2/3 - y/2} = \frac{6x(2 - x - y)}{4 - 3y} \text{ for } 0 < x < 1, 0 < y < 1. \end{aligned}$$

■

Example 5.8 ▷ *Joint Distribution: Radiators* ◁ A local radiator manufacturer subjects his radiators to two tests. The function that describes the percentage of radiators that pass the two tests is

$$f_{X,Y}(x, y) = 8xy, \quad 0 \leq y \leq x \leq 1 \quad (5.11)$$

The random variable X represents the percentage of radiators that pass test A , and Y represents the percentage of radiators that pass test B .

- (a) Is the function given in (5.11) a **pdf**?
- (b) Determine the marginal and conditional **pdfs** for X and Y .
- (c) Are X and Y independent?
- (d) Compute the probability that less than $\frac{1}{8}$ of the radiators will pass test B given that $\frac{1}{2}$ have passed test A .
- (e) Compute the quantities: $E[X]$, $E[X^2]$, $Var(X)$, $E[Y]$, $E[Y^2]$, and $Var(Y)$.
- (f) Use **S** to represent graphically (5.11).

Solution: The answers are as follows:

- (a) The function (5.11) is a **pdf** since $f_{X,Y}(x,y)$ is non-negative and

$$\int_0^1 \int_0^x 8xy \, dy \, dx = 8 \int_0^1 \left[x \int_0^x y \, dy \right] dx = 8 \int_0^1 \frac{x^3}{2} dx = 1$$

- (b) The marginal and conditional **pdfs** are

$$\begin{aligned} f_X(x) &= \int f(x,y) dy = \int_0^x 8xy \, dy = 4x^3, \quad 0 \leq x \leq 1 \\ f_Y(y) &= \int f(x,y) dx = \int_y^1 8xy \, dx = 4y(1-y^2), \quad 0 \leq y \leq 1 \\ f_{X|Y}(x|y) &= \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{8xy}{4y(1-y^2)} = \frac{2x}{1-y^2}, \quad y \leq x \leq 1 \\ f_{Y|X}(y|x) &= \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{8xy}{4x^3} = \frac{2y}{x^2}, \quad 0 \leq y \leq x \end{aligned}$$

- (c) The random variables X and Y are dependent since $f_{X,Y}(x,y) = 8xy \neq f_X(x) \cdot f_Y(y) = 16x^3y - 16x^3y^3$.

- (d) The probability that $\mathbb{P}(Y < 1/8 \mid X = 1/2)$ is computed as

$$\mathbb{P}(Y < 1/8 \mid X = 1/2) = \int_0^{1/8} f_{X,Y}(y|1/2) \, dy = \int_0^{1/8} \frac{2y}{1/4} \, dy = 4y^2 \Big|_0^{1/8} = \frac{1}{16}$$

(e) The quantities $E[X]$, $E[X^2]$, $\text{Var}(X)$, $E[Y]$, $E[Y^2]$, and $\text{Var}(Y)$ are

$$E[X] = \int_0^1 x \cdot 4x^3 dx = 4 \int_0^1 x^4 dx = \frac{4}{5}$$

$$E[X^2] = \int_0^1 x^2 \cdot 4x^3 dx = 4 \int_0^1 x^5 dx = \frac{2}{3}$$

$$\text{Var}(X) = E[X^2] - [E[X]]^2 = \frac{2}{3} - \frac{16}{25} = \frac{2}{75}$$

$$E[Y] = \int_0^1 y \cdot 4y(1-y^2) dy = 4 \int_0^1 (y^2 - y^4) dy = \frac{8}{15}$$

$$E[Y^2] = \int_0^1 y^2 \cdot 4y(1-y^2) dy = 4 \int_0^1 (y^3 - y^5) dy = \frac{1}{3}$$

$$\text{Var}(Y) = E[Y^2] - [E[Y]]^2 = \frac{1}{3} - \frac{64}{225} = \frac{11}{225}$$

(f) The following code can be used to create a graph similar to Figure 5.2:

```
> function.draw <- function(f, low=-1, hi=1, n=30){
+   r <- seq(low, hi, length=n)
+   z <- outer(r, r, f)
+   persp(r, r, z, xlab="X", ylab="Y", zlab="Z")}
> f3 <- function(x, y) {ifelse(x >= y, 8*x*y, 0)}
> function.draw(f3,0,1,25)
```

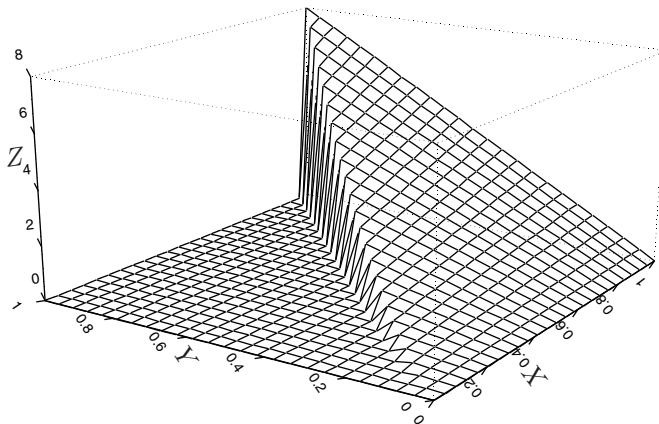


FIGURE 5.2: Graphical representation of $f_{X,Y}(x,y) = 8xy$, $0 \leq y \leq x \leq 1$ ■

Be careful not to assume the variance of the sum of two random variables is the sum of the variances of each random variable. Only if X and Y are independent is it true that $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$. A simple example to show why this is not true in general is computing $\text{Var}[X + X] \neq \text{Var}[X] + \text{Var}[X]$ since $\text{Var}[X + X] = \text{Var}[2X] = 4 \text{Var}[X]$. However, if X_1, X_2, \dots, X_n are n independent random variables with means $\mu_1, \mu_2, \dots, \mu_n$, and variances $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$, respectively, then the mean and variance of $Y = \sum_{i=1}^n c_i X_i$ where the c_i s are real-valued constants are $\mu_Y = \sum_{i=1}^n c_i \mu_i$ and $\sigma_Y^2 = \sum_{i=1}^n c_i^2 \sigma_i^2$. The proofs of the last two statements are left as exercises for the reader. (See problem 36 on page 196.)

Example 5.9 Let X_1, X_2, \dots, X_n be a random sample from a distribution with mean μ and standard deviation σ . Find the mean and variance of $Y = \frac{X_1 + X_2 + \dots + X_n}{n}$.

Solution: In the expression $Y = \frac{X_1 + X_2 + \dots + X_n}{n}$, the c_i values are all $\frac{1}{n}$. Consequently, $\mu_Y = \sum_{i=1}^n \frac{1}{n} \cdot \mu = \mu$ and $\sigma_Y^2 = \sum_{i=1}^n \left(\frac{1}{n}\right)^2 \cdot \sigma^2 = \frac{\sigma^2}{n}$. ■

5.5 Expected Values, Covariance, and Correlation

5.5.1 Expected Values

In Sections 3.4.3 on page 90 and 3.4.5.3 on page 98, the expected value for a single random variable for the discrete and continuous cases, respectively, was discussed. Also discussed was the expected value of a function of a random variable. In this section, the expected value of a function of two random variables is examined. When X and Y are jointly distributed random variables with **pdfs** $p_{X,Y}(x, y)$ or $f_{X,Y}(x, y)$, depending on whether the random variables are discrete or continuous, respectively, the expected value of $g(X, Y)$ is

$$E[g(X, Y)] = \begin{cases} \sum_x \sum_y g(x, y) \cdot p_{X,Y}(x, y) & \text{if } X \text{ and } Y \text{ are discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) \cdot f_{X,Y}(x, y) dx dy & \text{if } X \text{ and } Y \text{ are continuous} \end{cases} \quad (5.12)$$

The conditional expectation of X given a value y of Y is written

$$E[X|Y] = \begin{cases} \sum_x x \cdot p_{X|Y}(x|y) & \text{if } X \text{ and } Y \text{ are discrete} \\ \int_{-\infty}^{\infty} x \cdot f_{X|Y}(x|y) dx & \text{if } X \text{ and } Y \text{ are continuous} \end{cases} \quad (5.13)$$

Example 5.10 Let the random variables X and Y have a joint **pdf**:

$$f_{X,Y}(x, y) = \frac{e^{-y/x} e^{-x}}{x} \quad x > 0, \quad y > 0$$

Compute $E[Y|X = x]$.

Solution: First, compute the conditional **pdf** $f_{Y|X}(y|x)$:

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{f(x, y)}{f_X(x)} = \frac{\frac{e^{-y/x} e^{-x}}{x}}{\int_0^{\infty} \frac{e^{-y/x} e^{-x}}{x} dy} = \frac{\frac{e^{-y/x}}{x}}{\int_0^{\infty} \frac{e^{-y/x}}{x} dy} \\ &= \frac{e^{-y/x}}{x}, \quad x > 0, \quad y > 0 \end{aligned}$$

Using (5.13) for continuous random variables, write

$$E[Y|X = x] = \int_0^{\infty} y \cdot \frac{e^{-y/x}}{x} dy$$

Integrating by parts with $u = y$ and $dv = \frac{e^{-y/x}}{x}$, obtain

$$E[Y|X = x] = -ye^{-y/x} \Big|_0^{\infty} + \int_0^{\infty} e^{-y/x} dy = 0 + -xe^{-y/x} \Big|_0^{\infty} = x, \quad x > 0$$

When two random variables, say X and Y , are independent, recall that $f(x, y) = f_X(x) \cdot f_Y(y)$ for the continuous case and $p_{X,Y}(x, y) = p_X(x) \cdot p_Y(y)$ for the discrete case. Further, $E[XY] = E[X] \cdot E[Y]$. The last statement is true for both continuous and discrete X and Y . A proof for the discrete case is provided. Note that the proof in the continuous case would simply consist of exchanging the summation signs for integral signs.

Proof:

$$\begin{aligned} E[XY] &= \sum_x \sum_y xy p_{X,Y}(x, y) = \sum_x \sum_y xy p_X(x) p_Y(y) \\ &= \sum_y y p_Y(y) \sum_x x p_X(x) = E[Y]E[X] \end{aligned}$$

Example 5.11 Use the joint **pdf** provided in Example 5.8 on page 177 and compute $E[XY]$.

Solution:

$$E[XY] = \int_0^1 \int_0^x xy \cdot 8xy dy dx = 8 \int_0^1 \left[x^2 \int_0^x y^2 dy \right] dx = 8 \int_0^1 \frac{x^5}{3} dx = \frac{4}{9}$$

Since the random variables X and Y were found to be dependent in part (c) of Example 5.8 on page 177, note that

$$E[XY] = \frac{4}{9} \neq E[X] \cdot E[Y] = \frac{4}{5} \cdot \frac{8}{15} = \frac{32}{75}$$

5.5.2 Covariance

When two variables, X and Y , are not independent or when it is noted that $E[XY] \neq E[X] \cdot E[Y]$, one is naturally interested in some measure of their dependency. The covariance of X and Y , written $Cov[X, Y]$, provides one measure of the degree to which X and Y tend to move linearly in either the same or opposite directions. The covariance of two random variables X and Y is defined as

$$\begin{aligned} Cov[X, Y] &= E[(X - \mu_X)(Y - \mu_Y)] \\ &= \begin{cases} \sum_x \sum_y (x - \mu_X)(y - \mu_Y) p_{X,Y}(x, y) & X, Y \text{ discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f(x, y) dx dy & X, Y \text{ continuous} \end{cases} \end{aligned} \tag{5.14}$$

A $Cov[X, Y] > 0$ indicates that, generally, as X increases, so does Y (that is, X and Y move in the same direction); whereas, a $Cov[X, Y] < 0$ indicates that, generally, as X increases Y decreases (that is, X and Y move in opposite directions). To gain an intuitive understanding of covariance, see Figure 5.3, which has both horizontal and vertical dotted lines to indicate μ_{X_i} and μ_{Y_i} in each of the three plots. The first plot in Figure 5.3 exhibits a strong positive relationship. By this it is meant that large values of X tend to occur with large values of Y and small values of X tend to occur with small values of Y . Consequently, $(x - \mu_X)$ will tend to have the same sign as $(y - \mu_Y)$, so their product will be positive. In the center plot of Figure 5.3, the relationship between the two variables is negative, and note that $(x - \mu_{X_2})$ and $(y - \mu_{Y_2})$ tend to have opposite signs, which makes most of their products negative.

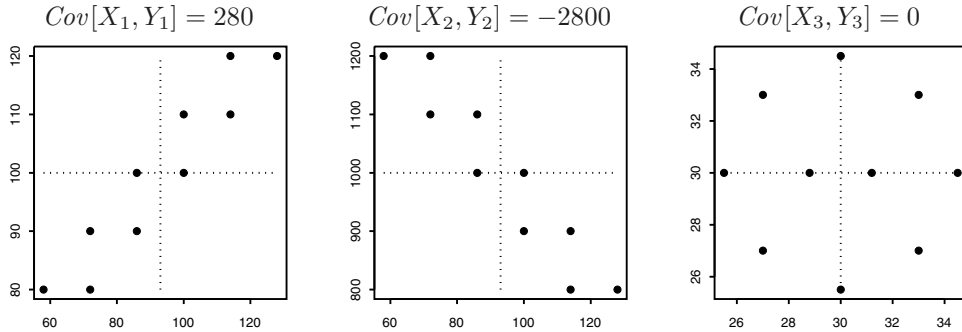


FIGURE 5.3: Scatterplots showing positive, negative, and zero covariance between two random variables where $p_{X,Y}(x, y) = \frac{1}{10}$ for each of the ten pairs of plotted points.

Example 5.12 Compute the covariance between X_1 and Y_1 for the values provided in Table 5.2 given that $p_{X,Y}(x, y) = \frac{1}{10}$ for each (x, y) pair.

Table 5.2: Values used to compute covariance for Figure 5.3

X_1	Y_1	X_2	Y_2	X_3	Y_3
58	80	58	1200	25.5	30.0
72	80	72	1200	27.0	33.0
72	90	72	1100	30.0	34.5
86	90	86	1100	33.0	33.0
86	100	86	1000	34.5	30.0
100	100	100	1000	33.0	27.0
100	110	100	900	30.0	25.5
114	110	114	900	27.0	27.0
114	120	114	800	28.8	30.0
128	120	128	800	31.2	30.0

Solution:

$$\begin{aligned}
 p_{X_1}(x) &= \sum_y p_{X_1, Y_1}(x, y) \\
 \mu_{X_1} &= \sum_x x \cdot p_{X_1}(x) = \frac{58 + 72 + \cdots + 128}{10} = 93 \\
 \mu_{Y_1} &= \sum_y y \cdot p_{Y_1}(y) = \frac{80 + 80 + \cdots + 120}{10} = 100 \\
 \text{Cov}[X_1, Y_1] &= \sum_x \sum_y (x - \mu_{X_1})(y - \mu_{Y_1})p_{X_1, Y_1}(x, y) \\
 &= (58 - 93) \cdot (80 - 100) \cdot \frac{1}{10} + (72 - 93) \cdot (80 - 100) \cdot \frac{1}{10} + \cdots \\
 &\quad + (128 - 120) \cdot (120 - 100) \cdot \frac{1}{10} \\
 &= 280
 \end{aligned}$$

To reduce the arithmetic drudgery, one can solve the problem with S:

```

> X1 <- c(58,72,72,86,86,100,100,114,114,128)
> Y1 <- c(80,80,90,90,100,100,110,110,120,120)
> covar <- function(x, y, f){sum((x-mean(x))*(y-mean(y))*f)}
> covar(X1, Y1,1/10)
[1] 280

```

There is a covariance function in R, however, it uses an unbiased estimator ($n - 1$) in the denominator instead of n . The covariance can be obtained directly with S-PLUS using the command `var(X, Y, unbiased=F)`. The S-PLUS command `var(X, Y, unbiased=T)` returns the same value as the R command `cov(X, Y)`. ■

At times, it will be easier to work with the shortcut formula $\text{Cov}[X, Y] = E[XY] - \mu_X \cdot \mu_Y$ instead of using the definition in (5.14).

Example 5.13 Compute the covariance between X and Y for Example 5.8 on page 177. In part (e) of Example 5.8, $E[X]$ and $E[Y]$ were computed to be $\frac{4}{5}$ and $\frac{8}{15}$, respectively, and in Example 5.11 on page 181, it was found that $E[XY] = \frac{4}{9}$.

Solution: Using the shortcut formula,

$$\text{Cov}[X, Y] = E[XY] - \mu_X \mu_Y = \frac{4}{9} - \frac{4}{5} \cdot \frac{8}{15} = \frac{4}{225}. \quad \blacksquare$$

When one examines the first two plots in Figure 5.3 on the preceding page, the dependency in the left plot seems to be about as strong as the dependency in the center plot, just in the opposite direction. However, the $\text{Cov}[X, Y] = 280$ in the left plot and $\text{Cov}[X, Y] = -2800$ in the center plot. It turns out that the dependencies are the same (just in opposite directions), but the units of measurement for the Y variable in the center plot are a factor of 10 times larger than those in the left plot. So, it turns out that covariance is unit dependent. To eliminate this unit dependency, scale the covariance.

5.5.3 Correlation

The **correlation coefficient** between X and Y , denoted $\rho_{X, Y}$, or simply ρ , is a scale-independent measure of linear dependency between two random variables. The independence

in scale is achieved by dividing the covariance by $\sigma_X\sigma_Y$. Specifically, define the correlation between X and Y as

$$\rho_{X,Y} = \frac{\text{Cov}[X, Y]}{\sigma_X\sigma_Y} \quad (5.15)$$

The correlation coefficient measures the degree of linear dependency between two random variables and is bounded by -1 and $+1$. The values $\rho = -1$ and $\rho = +1$ indicate perfect negative and positive relationships between two random variables. When $\rho = 0$, there is an absence of linear dependency between X and Y . If X and Y are independent, it is also true that $\rho = 0$; however, $\rho = 0$ does not imply independence. A similar statement is true for the $\text{Cov}[X, Y]$. That is, if X and Y are independent, $\text{Cov}[X, Y] = 0$; however, $\text{Cov}[X, Y] = 0$ does not imply independence.

Example 5.14 Compute $\rho_{X,Y}$ for Example 5.8 on page 177. Recall that $\text{Cov}[X, Y] = \frac{4}{225}$ was computed in Example 5.13 on the previous page, and $\text{Var}[X] = \frac{2}{75}$ and $\text{Var}[Y] = \frac{11}{225}$ in part (e) of Example 5.8 on page 177.

Solution:

$$\rho_{X,Y} = \frac{\text{Cov}[X, Y]}{\sigma_X\sigma_Y} = \frac{\frac{4}{225}}{\sqrt{\frac{2}{75} \cdot \frac{11}{225}}} = 0.4924 \quad \blacksquare$$

Example 5.15 Given the random variables X and Y with their joint probability distribution provided in Table 5.3, verify that although $\text{Cov}[X, Y] = 0$, X and Y are dependent.

Table 5.3: Joint probability distribution for X and Y

		Y		
		-1	0	1
X	-1	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
	0	$\frac{1}{8}$	0	$\frac{1}{8}$
	1	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

Solution: Start by computing the quantities $E[XY]$, $E[X]$, and $E[Y]$ to use in the shortcut formula for the covariance:

$$\begin{aligned} E[X] &= (-1) \cdot \frac{3}{8} + (0) \cdot \frac{2}{8} + (1) \cdot \frac{3}{8} = 0 \\ E[Y] &= (-1) \cdot \frac{3}{8} + (0) \cdot \frac{2}{8} + (1) \cdot \frac{3}{8} = 0 \\ E[XY] &= (-1 \cdot -1) \cdot \frac{1}{8} + \cdots + (1 \cdot 1) \cdot \frac{1}{8} = 0 \\ \text{Cov}[X, Y] &= E[XY] - E[X] \cdot E[Y] = 0 \end{aligned}$$

The covariance for this problem is 0. However, the random variables are dependent since

$$\mathbb{P}(X = -1, Y = -1) = \frac{1}{8} \neq \mathbb{P}(X = -1) \cdot \mathbb{P}(Y = -1) = \frac{3}{8} \cdot \frac{3}{8} = \frac{9}{64}.$$

This example reinforces the idea that a covariance or correlation coefficient of 0 does not imply independence. \blacksquare

Example 5.16 Compute ρ_{X_1, Y_1} for Example 5.12 on page 182. Recall that $\mu_{X_1} = 93$, $\mu_{Y_1} = 100$, and $Cov[X_1, Y_1] = 280$.

Solution: Start by computing the quantities $E[X_1^2]$, $E[Y_1^2]$, σ_{X_1} , and σ_{X_2} :

$$\begin{aligned} E[X_1^2] &= \sum_x x^2 p_{X_1}(x) \\ &= 58^2 \cdot \frac{1}{10} + 72^2 \cdot \frac{1}{10} + \cdots + 128^2 \cdot \frac{1}{10} = 9090 \\ E[Y_1^2] &= \sum_y y^2 p_{Y_1}(y) \\ &= 80^2 \cdot \frac{1}{10} + 80^2 \cdot \frac{1}{10} + \cdots + 120^2 \cdot \frac{1}{10} = 10200 \\ Var[X_1] &= E[X_1^2] - (E[X_1])^2 = 9090 - 93^2 = 441 \\ \sigma_{X_1} &= \sqrt{Var[X_1]} = \sqrt{441} = 21 \\ Var[Y_1] &= E[Y_1^2] - (E[Y_1])^2 = 10200 - 100^2 = 200 \\ \sigma_{Y_1} &= \sqrt{Var[Y_1]} = \sqrt{200} = 14.14214 \\ \rho_{X_1, Y_1} &= \frac{Cov[X_1, Y_1]}{\sigma_{X_1} \sigma_{Y_1}} = \frac{280}{21 \times 14.14214} = 0.9428087 \end{aligned}$$

It is also possible to get the answer directly from S by entering

```
> cor(X1, Y1)
[1] 0.942809
```

■

It is worthwhile to note that $\rho_{X_1, Y_1} = 0.9428087$ and $\rho_{X_2, Y_2} = -0.9428087$ for the left and center plots, respectively, in Figure 5.3 on page 182. In other words, the correlations have the same absolute magnitude for both plots, even though the absolute values of the covariances differ by a factor of ten.

5.6 Multinomial Distribution

The multinomial distribution is a generalization of the binomial distribution. Recall that each trial in a binomial experiment results in only one of two mutually exclusive outcomes. Experiments where each trial can result in any one of k possible mutually exclusive outcomes A_1, \dots, A_k with probabilities $\mathbb{P}(A_i) = \pi_i$, $0 < \pi_i < 1$, for $i = 1, \dots, k$ such that $\sum_{i=1}^k \pi_i = 1$ can be modeled with the **multinomial distribution**. Specifically, the multinomial distribution computes the probability that A_1 occurs x_1 times, A_2 occurs x_2 times, \dots , A_k occurs x_k times in n independent trials, where $x_1 + x_2 + \cdots + x_k = n$. To derive the probability distribution function, reason in a fashion similar to that done with the binomial. Since the trials are independent, any specified ordering yielding x_1 outcomes for A_1 , x_2 outcomes for A_2 , \dots , and x_k outcomes for A_k will occur with probability $\pi_1^{x_1} \pi_2^{x_2} \cdots \pi_k^{x_k}$. The total number of orderings yielding x_1 outcomes for A_1 , x_2 outcomes for A_2 , \dots , and x_k outcomes for A_k is $\frac{n!}{x_1! x_2! \cdots x_k!}$. With these two facts in mind, the probability distribution, mean, variance, and **mgf** of a multinomial distribution can be derived. All are found in (5.16).

Multinomial Distribution

$$\mathbf{X} \sim MN(n, \pi_1, \dots, \pi_k)$$

$$\mathbb{P}(\mathbf{X} = (x_1, \dots, x_k) | n, \pi_1, \dots, \pi_k) = \frac{n!}{x_1! x_2! \dots x_k!} \pi_1^{x_1} \pi_2^{x_2} \dots \pi_k^{x_k} \quad (5.16)$$

$$E[X_i] = n\pi_i$$

$$\text{Var}[X_i] = n\pi_i(1 - \pi_i)$$

given that each $X_i \sim \text{Bin}(n, \pi_i)$

$$M_{\mathbf{X}}(t) = (\pi_1 e^{t_1} + \pi_2 e^{t_2} + \dots + \pi_{k-1} e^{t_{k-1}} + \pi_k e^{t_k})^n$$

Example 5.17 The probability a particular type of light bulb lasts less than 500 hours is 0.5 and the probability the same type of light bulb lasts more than 800 hours is 0.2. In a random sample of ten light bulbs, what is the probability of obtaining exactly four light bulbs that last less than 500 hours and two light bulbs that last more than 800 hours?

Solution: Let the random variables X_1 , X_2 , and X_3 denote the number of light bulbs that last less than 500 hours, the number of light bulbs that last between 500 and 800 hours, and the number of light bulbs that last more than 800 hours, respectively. Since $\pi_1 = 0.5$, $\pi_2 = 0.3$, and $\pi_3 = 0.2$, use the first equation in (5.16) and compute $\mathbb{P}(X_1 = 4, X_2 = 4, X_3 = 2)$ as

$$\mathbb{P}(X_1 = 4, X_2 = 4, X_3 = 2 | 10, 0.5, 0.3, 0.2) = \frac{10!}{4!4!2!} (0.5)^4 (0.3)^4 (0.2)^2 = 0.0638. \quad \blacksquare$$

5.7 Bivariate Normal Distribution

The joint distribution of the random variables X and Y is said to have a **bivariate normal** distribution when its joint density takes the form

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)} \left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho \left(\frac{x-\mu_X}{\sigma_X}\right) \left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 \right]\right\}, \quad (5.17)$$

for $-\infty < x, y < +\infty$, where $\mu_X = E[X]$, $\mu_Y = E[Y]$, $\sigma_X^2 = \text{Var}[X]$, $\sigma_Y^2 = \text{Var}[Y]$, and ρ is the correlation coefficient between X and Y . An equivalent representation of (5.17) is given in (5.18), where $\mathbf{X} = (X, Y)^T$ is a vector of random variables where T represents the transpose, $\boldsymbol{\mu} = (\mu_X, \mu_Y)^T$, is a vector of constants, and $\boldsymbol{\Sigma}$ is a 2×2 non-singular matrix such that its inverse $\boldsymbol{\Sigma}^{-1}$ exists and the determinant $|\boldsymbol{\Sigma}| \neq 0$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} \text{Var}[X] & \text{Cov}[X, Y] \\ \text{Cov}[Y, X] & \text{Var}[Y] \end{pmatrix}.$$

$$f(\mathbf{x}) = \frac{1}{(\sqrt{2\pi})^2} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{X} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) \right\}. \quad (5.18)$$

The shorthand notation used to denote a multivariate (bivariate being a subset) normal distribution is $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$. In general, Σ represents what is called the variance covariance matrix. When $\mathbf{X} = (X_1, X_2, \dots, X_n)^T$ and $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_n)^T$ it is defined as

$$\begin{aligned} \Sigma &= E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T] = E \left[\begin{pmatrix} X_1 - \mu_1 \\ \vdots \\ X_n - \mu_n \end{pmatrix} (X_1 - \mu_1, \dots, X_n - \mu_n) \right] \\ &= \begin{pmatrix} \sigma_{X_1}^2 & \dots & Cov(X_1, X_n) \\ \vdots & \ddots & \vdots \\ Cov(X_n, X_1) & \dots & \sigma_{X_n}^2 \end{pmatrix}. \end{aligned}$$

Different representations of four bivariate normal distributions, all with parameters $\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_Y = 1$, and ρ values of 0, 0.30, 0.60, and 0.95, respectively, are provided in Figure 5.4 on the following page. The following code produces a perspective plot, contour plot, and image plot of a bivariate normal density with parameters $\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_Y = 1$, and $\rho = 0.5$ similar to those in Figure 5.4 on the next page:

```
> function1.draw <- function(f, low = -1, hi = 1, n = 50){
+   r <- seq(low, hi, length = n)
+   z <- outer(r, r, f)
+   persp(r, r, z, axes=FALSE, box=TRUE)}
> par(mfrow=c(1,3), pty="s")
> f1 <- function(x, y){
+   exp( (x^2-2*0.5*x*y+y^2) / (-2*(1-0.5^2)) ) /
+   (2*pi*sqrt(1-0.5^2))}
> x <- seq(-3,3, length=100)
> y <- x
> function1.draw(f1,-3,3,20)
> contour(x, y, outer(x, y, f1), nlevels=10)
> image(x, y, outer(x, y, f1), zlim=range(outer(x, y, f1)), add = FALSE)
```

The following facts about the bivariate normal distribution are listed without proof:

- (a) The marginal distribution of X is $N(\mu_X, \sigma_X)$.
- (b) The marginal distribution of Y is $N(\mu_Y, \sigma_Y)$.
- (c) If X and Y have a bivariate normal distribution, the conditional density of Y given $X = x$ is a normal distribution with mean $\mu_{Y|x} = E(Y|x) = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X} (x - \mu_X)$ and variance $\sigma_{Y|x}^2 = \sigma_Y^2 (1 - \rho^2)$.
- (d) Given any two constants a and b , the distribution of $aX + bY$ is

$$N \left(a\mu_X + b\mu_Y, \sqrt{a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\rho\sigma_X\sigma_Y} \right)$$

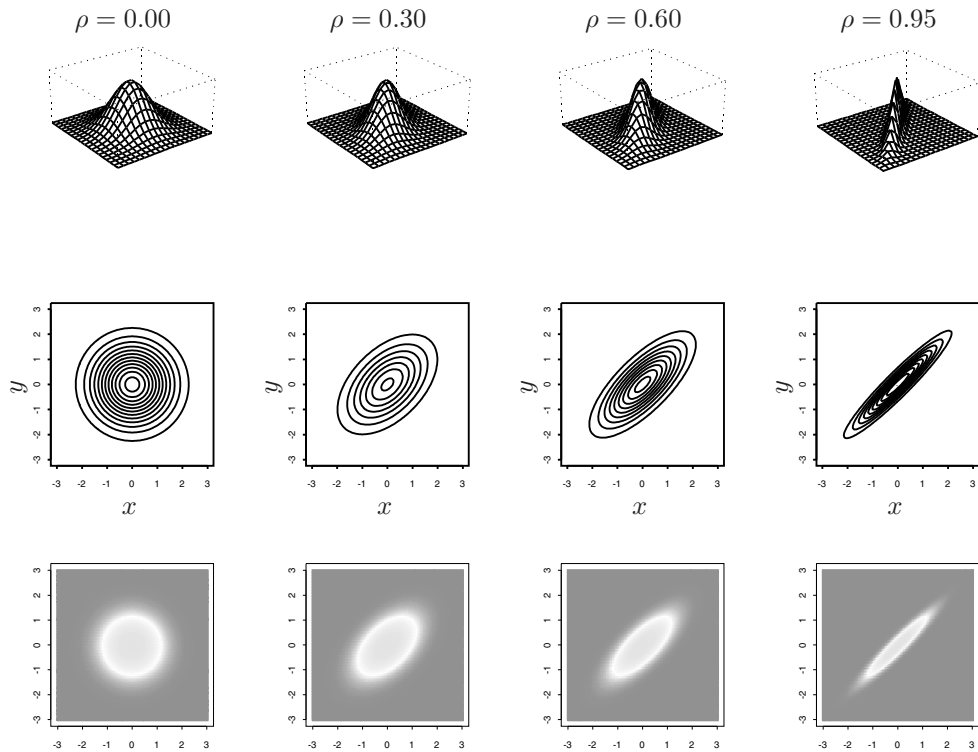


FIGURE 5.4: The first row uses the function perspective to represent bivariate normal densities with parameters $\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_Y = 1$, and ρ values of 0, 0.30, 0.60, and 0.95, respectively. The second row represents the same bivariate densities with contour plots, while the third row represents the densities with image plots.

Example 5.18 ▷ *Bivariate Normal Grades* ◁ Let us assume that the distribution of grades for a particular group of students where X and Y represent the grade point averages in high school and the first year of college, respectively, follow a bivariate normal distribution with parameters $\mu_X = 3.2$, $\mu_Y = 2.4$, $\sigma_X = 0.4$, $\sigma_Y = 0.6$, and $\rho = 0.6$. Find the following:

- $\mathbb{P}(Y < 1.8)$
- $\mathbb{P}(Y < 1.8 | X = 2.5)$
- $\mathbb{P}(Y > 3.0)$
- $\mathbb{P}(Y > 3.0 | X = 2.5)$

Solution: The answers are computed first manually, and then with S.

- Using the parameters given in the problem,

$$\mathbb{P}(Y < 1.8) = \mathbb{P}\left(\frac{Y - 2.4}{0.6} < \frac{1.8 - 2.4}{0.6}\right) = \mathbb{P}(Z < -1) = 0.1586$$

```
> pnorm(1.8, 2.4, .6)
[1] 0.1586553
```

(b) First, find the quantities $\mu_{Y|x=2.5}$ and $\sigma_{Y|x=2.5}$:

$$\mu_{Y|x=2.5} = E(Y|x = 2.5) = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X) = 2.4 + 0.6 \cdot \frac{0.6}{0.4} \cdot (2.5 - 3.2) = 1.77$$

$$\sigma_{Y|x=2.5}^2 = \sigma_Y^2(1 - \rho^2) = 0.6^2 \cdot (1 - 0.6^2) = 0.2304 \Rightarrow \sigma_{Y|x=2.5} = 0.48$$

$$\mathbb{P}(Y < 1.8 | X = 2.5) = \mathbb{P}\left(\frac{Y - 1.77}{0.48} < \frac{1.8 - 1.77}{0.48}\right) = \mathbb{P}(Z < 0.0625) = 0.5249.$$

```
> pnorm(1.8, 1.77, .48)
[1] 0.5249177
```

(c) Using the parameters given in the problem,

$$\begin{aligned} \mathbb{P}(Y > 3.0) &= 1 - \mathbb{P}(Y \leq 3.0) = 1 - \mathbb{P}\left(\frac{Y - 2.4}{0.6} \leq \frac{3.0 - 2.4}{0.6}\right) \\ &= 1 - \mathbb{P}(Z \leq 1) = 0.1586 \end{aligned}$$

```
> 1-pnorm(3, 2.4, .6)
[1] 0.1586553
```

(d) Using the quantities $\mu_{Y|x}$ and $\sigma_{Y|x}$ from (b),

$$\begin{aligned} \mathbb{P}(Y > 3.0 | X = 2.5) &= 1 - \mathbb{P}(Y \leq 3.0 | X = 2.5) \\ &= 1 - \mathbb{P}\left(\frac{Y - 1.77}{0.48} \leq \frac{3.0 - 1.77}{0.48}\right) = 1 - \mathbb{P}(Z \leq 2.5625) \\ &= 0.0052. \end{aligned}$$

```
> 1-pnorm(3, 1.77, .48)
[1] 0.005196079
```



5.8 Problems

1. Let X and Y have the following joint distribution:

Joint Probability Distribution of X and Y

		Y		
		-1	0	1
X	-1	1/6	0	1/6
	0	1/3	0	0
	1	1/6	0	1/6

- (a) Find the covariance between X and Y .
 (b) Show that X and Y are dependent.
2. Given the random variables X and Y and their joint probability $p_{X,Y}(X, Y)$:

		Y		
		1	2	3
X	1	0.05	0.05	0.1
	2	0.05	0.1	0.35
	3	0	0.2	0.1

- (a) Show that $p_{X,Y}(X, Y)$ satisfies properties (i) and (ii) given on page 171 for the joint **pdf** of two discrete random variables.
 (b) Find the mean of X and the mean of Y .
 (c) Are X and Y independent?
 (d) Find the variances of X and of Y .
 (e) Find the covariance of X and Y .
3. A particular unfair coin is constructed so that the probability of obtaining a head is $1/3$. The unfair coin is flipped twice. Define two random variables: Z = the number of heads in the first flip and W = the number of heads in two flips.
- (a) Construct a table showing the joint probability distribution of both random variables Z and W including the marginal probabilities.
 (b) Find the covariance between Z and W . Are they independent?
 (c) Suppose the covariance between Z and W were 0. Would this imply that Z and W are independent?
4. An international travel agency translates its promotional fliers each season. Translators are hired to translate the fliers into several languages. The translators are paid either € 60 or €90 per page, depending on word density. The fliers are all either 5, 7, or 10 pages in length. The joint density function for X and Y , where X = number of pages and Y = price per page, is

		Y	
		60	90
X	5	0.05	0.4
	7	0.05	0.1
	10	0.35	0.05

- (a) Find the mean and variance of X and Y .
- (b) Find $Cov(X, Y)$, and explain its meaning.
- (c) Find the probability function of Z (the total translation cost).
- (d) Find the mean of Z .
5. A student uses a free dialup service to access the Internet. Depending on the server to which the Internet service provider connects the student, there are three transmission rates: 1800, 2700, and 3600 bytes per second. Let X be the number of transmitted bytes and Y the transmission rate in bytes per second. The joint probability for X and Y is given by the following table:

		Y		
		1800	2700	3600
X	64800	0.3	0.05	0.025
	324000	0.025	0.15	0.15
	972000	0	0.2	0.1

- (a) Let Z be the random variable indicating the time necessary for transmission. Write down the probability function of Z .
- (b) Find the expected time spent in transmission.
- (c) Find the mean and variance of X and Y and $Cov(X, Y)$.
6. At the local movie theater, drinks and popcorn come in three sizes: small, medium, and large. The prices for both drinks and popcorn are \$1.50, \$2.50, and \$3.50 for the small, medium, and large sizes, respectively. For a given customer, define the random variables X = amount spent for popcorn and Y = amount spent for drinks. Suppose the joint distribution for X and Y is

		X		
		1.5	2.5	3.5
Y	1.5	0.03	0.07	0.05
	2.5	0.08	0.08	0.30
	3.5	0.00	0.30	0.09

- (a) Find the probability a given customer spends no more than \$2.50 on popcorn. What is the probability a given customer spends at least \$2.50 on popcorn?
- (b) What is the average amount of money spent at the movies for a customer buying both popcorn and a drink, if the cost of the movie ticket is \$5.20?
7. The interior diameter of a particular type of test tube is a random variable with a mean of 5 cm and a standard deviation of 0.03 cm. If the test tube thickness is a random variable with a mean of 0.5 cm and a standard deviation of 0.001 cm and both variables are independent, find the mean and standard deviation of the exterior diameter.

8. The flow of water arriving at an irrigation canal is measured in cubic meters and follows a $N(100, 20)$ distribution. The canal has a flow capacity that follows a $N(120, 30)$ distribution. The sluice gate is opened when the water flow exceeds the canal's capacity. What is the probability that the flood gate will be opened?
9. Given the joint density function

$$f(x, y) = 6x, \quad 0 < x < y < 1,$$

find the $E[Y | X]$ that is the regression line resulting from regressing Y on X .

10. The time, in minutes, that a car is parked in a mall has the following density function:

$$f(x) = \begin{cases} \frac{1}{50}e^{-x/50} & x > 0 \\ 0 & x \leq 0. \end{cases}$$

Using S,

- (a) Find the probability that a car stays more than 1 hour.
- (b) Let $Y = 0.5 + 0.03X$ be the cost in dollars that the mall has to pay a security service per parked car. Find the mean parking cost for 1000 cars.
- (c) Find the variance and skewness coefficient of Y .
11. A poker hand (5 cards) is dealt from a single deck of well shuffled cards. If the random variables X and Y represent the number of aces and the number of kings in a hand, respectively,
- (a) Write the joint distribution $f_{X,Y}(x, y)$.
- (b) What is the marginal distribution of X , $f_X(x)$?
- (c) What is the marginal distribution of Y , $f_Y(y)$?

$$\left(\text{Hint: } \sum_{y=0}^{\infty} \binom{a}{x} \binom{b}{n-x} = \binom{a+b}{n} \right).$$

12. If $f_{X,Y}(x, y) = 5x - y^2$ in the region bounded by $y = 0$, $x = 0$, and $y = 2 - 2x$, find the density function for the marginal distribution of X , for $0 < x < 1$.
13. If $f(x, y) = e^{-(x+y)}$, $x > 0$, and $y > 0$, find $\mathbb{P}(X + 3 > Y | X > \frac{1}{3})$.
14. If $f(x, y) = 1$, $0 < x < 1$, $0 < y < 1$, what is $\mathbb{P}(Y - X > \frac{1}{2} | X + Y > \frac{1}{2})$?
15. If $f(x, y) = k(y - 2x)$ is a joint density function over $0 < x < 1$, $0 < y < 1$, and $y > x^2$, then what is the value of the constant k ?
16. Let X and Y have the joint density function

$$f(x, y) = \begin{cases} \frac{4}{3}x + \frac{2}{3}y & \text{for } 0 < x < 1 \text{ and } 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

Find $\mathbb{P}(2X < 1 | X + Y < 1)$.

17. Let X and Y have the joint density function

$$f(x, y) = \begin{cases} 6(x - y)^2 & \text{for } 0 < x < 1 \text{ and } 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

- (a) Find $\mathbb{P}(X < \frac{1}{2} \mid Y < \frac{1}{4})$.
- (b) Find $\mathbb{P}(X < \frac{1}{2} \mid Y = \frac{1}{4})$.
18. Let X and Y denote the weight (in kilograms) and height (in centimeters), respectively, of 20-year-old American males. Assume that X and Y have a bivariate normal distribution with parameters $\mu_X = 82$, $\sigma_X = 9$, $\mu_Y = 190$, $\sigma_Y = 10$, and $\rho = 0.8$. Find
- (a) $E[Y \mid X = 75]$,
- (b) $E[Y \mid X = 90]$,
- (c) $\text{Var}[Y \mid X = 75]$,
- (d) $\text{Var}[Y \mid X = 90]$,
- (e) $\mathbb{P}(Y \geq 190 \mid X = 75)$, and
- (f) $\mathbb{P}(185 \leq Y \leq 195 \mid X = 90)$.
19. Let X and Y denote the heart rate (in beats per minute) and average power output (in watts) for a 10 minute cycling time trial performed by a professional cyclist. Assume that X and Y have a bivariate normal distribution with parameters $\mu_X = 180$, $\sigma_X = 10$, $\mu_Y = 400$, $\sigma_Y = 50$, and $\rho = 0.9$. Find
- (a) $E[Y \mid X = 170]$,
- (b) $E[Y \mid X = 200]$,
- (c) $\text{Var}[Y \mid X = 170]$,
- (d) $\text{Var}[Y \mid X = 200]$,
- (e) $\mathbb{P}(Y \leq 380 \mid X = 170)$, and
- (f) $\mathbb{P}(Y \geq 450 \mid X = 200)$.
20. A certain group of college students takes both the Scholastic Aptitude Test (SAT) and an intelligence quotient (IQ) test. Let X and Y denote the students' scores on the SAT and IQ tests, respectively. Assume that X and Y have a bivariate normal distribution with parameters $\mu_X = 980$, $\sigma_X = 126$, $\mu_Y = 117$, $\sigma_Y = 7.2$, and $\rho = 0.58$. Find
- (a) $E[Y \mid X = 1350]$,
- (b) $E[Y \mid X = 700]$,
- (c) $\text{Var}[Y \mid X = 700]$,
- (d) $\mathbb{P}(Y \leq 120 \mid X = 1350)$, and
- (e) $\mathbb{P}(Y \geq 100 \mid X = 700)$.
21. A pepper canning company uses tins weighing 20 grams. The full tin of peppers is placed on a balance. Customer good will is maximized when the balance shows a quantity μ and the peppers weight is Y grams. If the balance has a random error $X \sim N(0, \sigma = 10)$,
- (a) Find the relationship between Y , X , and μ .
- (b) What is the distribution of Y ?

- (c) Find μ so that 98% of the tins have at least 400 grams of peppers.
- (d) Repeat the exercise assuming that the tin weight is a random variable $W \sim N(20, \sigma = 5)$.
22. Given the joint density function $f_{XY}(x, y) = x + y$, $x \geq 0, y \leq 1$,
- (a) Show that properties (1) and (2) on page 173 for the joint **pdf** of two continuous random variables are satisfied.
- (b) Find the cumulative distribution function.
- (c) Find the marginal means of X and Y .
- (d) Find the marginal variances of X and Y .

23. The lifetime of two electronic components are two random variables, X and Y . Their joint density function is given by

$$f_{XY}(x, y) = \frac{1 + x + y + cxy}{(c + 3)} \exp(-(x + y)) \quad x \geq 0 \text{ and } y \geq 0$$

- (a) Verify that $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) dx dy = 1$.
- (b) Find $f_X(x)$.
- (c) What value of c makes X and Y independent?
24. A high technology company manufactures circular mirrors used in certain satellites. The radius of any mirror in inches is a random variable R with density function

$$f(r) = \begin{cases} \frac{24}{11}(2r - r^2) & 1 \leq r \leq \frac{3}{2} \\ 0 & \text{otherwise.} \end{cases}$$

To place the mirrors in the satellites without any problems, the mirror area, given by πR^2 , cannot be greater than 6.5 inches. Using S,

- (a) Verify that $\int_{-\infty}^{\infty} f(r) dr = 1$.
- (b) Find the mean area of the mirrors.
- (c) Find the probability that a mirror's diameter does not surpass 6.5 inches.
25. Use the package `adapt` from R to solve Example 5.2 on page 173.
26. Let X and Y have the joint density function

$$f_{XY}(x, y) = \begin{cases} Kxy & 2 \leq x \leq 4 \text{ and } 4 \leq y \leq 6 \\ 0 & \text{otherwise} \end{cases}$$

- (a) Find K so that the given function is a valid **pdf**.
- (b) Find the marginal densities of X and Y .
- (c) Are X and Y independent? Justify.

27. Given the joint density function of X and Y

$$f_{XY}(x, y) = \begin{cases} 1/2 & x + y \leq 2, \quad x \geq 0, \quad y \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

- (a) Find the marginal densities of X and Y .
 (b) Find $E[X]$, $E[Y]$, $Cov[X, Y]$, and $\rho_{X, Y}$.
 (c) Find $P(X + Y < 1 \mid X > \frac{1}{2})$.

28. Let X and Y have the joint density function

$$f_{XY}(x, y) = \begin{cases} Ky & -2 \leq x \leq 2, \quad 1 \leq y \leq x^2 \\ 0 & \text{otherwise.} \end{cases}$$

- (a) Find K so that $f_{XY}(x, y)$ is a valid **pdf**.
 (b) Find the marginal densities of X and Y .
 (c) Find $P(Y > \frac{3}{2} \mid X < \frac{1}{2})$.

29. An engineer has designed a new diesel motor that is used in a prototype vehicle. The prototype's diesel consumption in gallons per mile C follows the equation $C = 3 + 2X + \frac{3}{2}Y$, where X is a speed coefficient and Y is the quality diesel coefficient. Suppose the joint density for X and Y is $f_{XY}(x, y) = ky$, $0 \leq x \leq 2$, $0 \leq y \leq x$.

- (a) Find k so that $f_{XY}(x, y)$ is a valid density function.
 (b) Are X and Y independent?
 (c) Find the mean and variance for the prototype vehicle's diesel consumption.

30. To make porcelain, kaolin X and feldspar Y are needed to create a soft mixture that later becomes hard. The proportion of these components for every tone of porcelain has the density function $f_{XY}(x, y) = Kx^2y$, $0 \leq x \leq y \leq 1$, $x + y \leq 1$.

- (a) Find the value of K so that $f_{XY}(x, y)$ is a valid **pdf**.
 (b) Find the marginal densities of X and Y .
 (c) Find the kaolin mean and the feldspar mean by tone.
 (d) Find the probability that the proportion of feldspar will be higher than $1/3$, if the kaolin is more than half of the porcelain.

31. A device can fail in four different ways with probabilities $\pi_1 = 0.2$, $\pi_2 = 0.1$, $\pi_3 = 0.4$, and $\pi_4 = 0.3$. Suppose there are 12 devices that fail independently of one another. What is the probability of 3 failures of the first kind, 4 of the second, 3 of the third, and 2 of the fourth?

32. The wait time in minutes a shopper spends in a local supermarket's checkout line has distribution $f(x) = \exp(-x/2)/2$, $x > 0$. On weekends, however, the wait is longer, and the distribution then is given by $g(x) = \exp(-x/3)/3$, $x > 0$. Find

- (a) The probability that the waiting time for a customer will be less than 1 minute.
 (b) The probability that, given a waiting time of 2 minutes, it will be a weekend.

- (c) The probability that the customer waits less than 2 minutes.
33. An engineering team has designed a lamp with two light bulbs. Let X be the lifetime for bulb 1 and Y the lifetime for bulb 2, both in thousands of hours. Suppose that X and Y are independent and they follow an $\exp(\lambda = 1)$ distribution.
- (a) Find the joint density function of X and Y . What is the probability neither bulb lasts longer than 1000 hours?
- (b) If the lamp works when at least one bulb is lit, what is the probability that the lamp works no more than 2000 hours?
- (c) What is the probability that the lamp works between 1000 and 2000 hours?
34. The national weather service has issued a severe weather advisory for a particular county that indicates that severe thunderstorms will occur between 9 p.m. and 10 p.m. When the rain starts, the county places a call to the maintenance supervisor who opens the sluice gate to avoid flooding. Assuming the rain's start time is uniformly distributed between 9 p.m. and 10 p.m.
- (a) At what time, on the average, will the county maintenance supervisor open the sluice gate?
- (b) What is the probability that the sluice gate will be opened before 9:30 p.m.?

Note: Solve this problem both by hand and using S.

35. Example 5.18 on page 188 assumes the distribution of grades for a particular group of students, where X and Y represent the grade point averages in high school and the first year of college, respectively, and have a bivariate normal distribution with parameters $\mu_X = 3.2$, $\mu_Y = 2.4$, $\sigma_X = 0.4$, $\sigma_Y = 0.6$, and $\rho = 0.6$.
- (a) Set the seed equal to 194 (`set.seed(194)`), and use the function `mvrnorm()` from the MASS package to simulate the population, assuming the population of interest consists of 200 students. (Hint: Use `empirical=TRUE`.)
- (b) Compute the means of X and Y . Are they equal to 3.2 and 2.4, respectively?
- (c) Compute the variance of X and Y as well as the covariance between X and Y . Are the values 0.16, 0.36, and 0.144, respectively?
- (d) Create a scatterplot of Y versus X . If a different seed value is used, how do the simulated numbers differ?
36. Show that if X_1, X_2, \dots, X_n are independent random variables with means $\mu_1, \mu_2, \dots, \mu_n$ and variances $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$, respectively, then the mean and variance of $Y = \sum_{i=1}^n c_i \mu_i$, where the c_i s are real-valued constants, are $\mu_Y = \sum_{i=1}^n c_i \mu_i$ and $\sigma_Y^2 = \sum_{i=1}^n c_i^2 \sigma_i^2$ as stated on page 180 of the text. (Hint: Use moment generating functions.)

Chapter 6

Sampling and Sampling Distributions

6.1 Sampling

The objective of statistical analysis is to gain knowledge about certain properties in a population that are of interest to the researcher. When the population is small, the best way to study the population of interest is to study all of the elements in the population one by one. This process of collecting information on the entire population of interest is called a **census**. However, it is usually quite challenging to collect information on an entire population of interest. Not only do monetary and time constraints prevent a census from being taken easily, but also the challenges of finding all the members of a population can make gathering an accurate census all but impossible. Under certain conditions, a random selection of certain elements actually returns more reliable information than can be obtained by using a census. Standard methods used to learn about the characteristics of a population of interest include simulation, designed experiments, and sampling.

Simulation studies typically generate numbers according to a researcher-specified model. For a simulation study to be successful, the chosen simulation model must closely follow the real life process the researcher is attempting to simulate. For example, the effects of natural disasters, such as earthquakes, on buildings and highways are often modeled with simulation.

When the researcher has the ability to control the research environment, or at least certain variables of interest in the study, **designed experiments** are typically employed. The objective of designed experiments is to gain an understanding about the influence that various levels of a factor have on the response of a given experiment. For example, an agricultural researcher may be interested in determining the optimal level of nitrogen when his company's fertilizer is used to grow wheat in a particular type of soil. The designed experiment might consist of applying the company's fertilizer to similar plots using three different concentrations of nitrogen in the fertilizer.

Sampling is the most frequently used form of collecting information about a population of interest. Many forms of sampling exist, such as random sampling, simple random sampling, systematic sampling, and cluster sampling. It will be assumed that the population from which one is sampling has size N and that the sample is of size $n < N$.

Random sampling is the process of selecting n elements from a population where each of the n elements has the same probability of being selected, namely, $\frac{1}{N}$. More precisely, the random variables X_1, X_2, \dots, X_n form a random sample of size n from a population with a **pdf** $f(x)$ if X_1, X_2, \dots, X_n are mutually independent random variables such that the marginal **pdf** of each X_i is $f(x)$. The statement " X_1, X_2, \dots, X_n are independent and identically distributed, i.i.d., random variables with **pdf** $f(x)$ " is often used to denote a random sample. The objective of random sampling is to obtain a representative sample of the population that can be used to make generalizations about the population.

This process of making generalizations about the population from sampled information

is called **inferential statistics**. For the generalizations to be valid, the sample must meet certain requirements. The key requirement for a random sample is that it be representative of the parent population from which it was taken.

The typical method of obtaining a random sample starts with using either a calculator or a computer random number generator to decide which elements of a population to sample. The numbers returned from random number generating functions are not, in the strictest sense, random. That is, because an algorithm is used to generate the numbers, they are not completely random. Depending on the quality or lack thereof for a given random number generator, the same numbers may begin to cycle after a number of iterations. This problem is encountered much less with the random number generating functions written for computers than it is with those for calculators. In general, random number generators return pseudo-random numbers from a $Unif(0,1)$ distribution. Since people tend to favor certain numbers, it is best not to allow humans to pick random numbers unless the process is one of selecting numbers from an urn or another similar process. To avoid possible biases, it is best to let a function written to generate random numbers pick a sample.

When the population is finite, it is possible to list all of the possible combinations of samples of size n using the S command `expand.grid()`. For example, suppose all of the combinations of size $n = 3$ from a population consisting of $N = 4$ items are to be listed. Clearly, there are $4 \times 4 \times 4 = 64$ possible combinations. To enumerate the possible combinations with S, type `expand.grid(1:4,1:4,1:4)`. In a similar fashion, if all of the possible combinations from rolling two fair dice or all possible combinations of size $n = 2$ from the population $X_1 = 2$, $X_2 = 5$, and $X_3 = 8$ are to be enumerated, type `expand.grid(1:6,1:6)` or `expand.grid(c(2,5,8), c(2,5,8))`, respectively.

6.1.1 Simple Random Sampling

Simple random sampling is the most elementary form of sampling. In a simple random sample, each particular sample of size n has the same probability of occurring. In finite populations, each of the $\binom{N}{n}$ samples of size n is taken without replacement and has the same probability of occurring. If the population being sampled is infinite, the distinction between sampling with replacement and sampling without replacement becomes moot. That is, in an infinite population, the probability of selecting a given element is the same whether sampling is done with or without replacement. Conceptually, the population can be thought of as balls in an urn, a fixed number of which are randomly selected without replacement for the sample. Most sampling is done without replacement due to its ease and increased efficiency in terms of variability compared to sampling with replacement.

To list all of the possible combinations of size n when sampling without replacement from a finite population of size N , that is, the $\binom{N}{n}$ combinations, the function `Combinations()` written by Tim Hesterberg at Insightful can be used. Make sure the PASWR package is loaded, as it contains the function `Combinations()`.

Example 6.1 Given a population of size $N = 5$, use S to list all of the possible samples of size $n = 3$. That is, list the $\binom{5}{3} = 10$ possible combinations.

Solution: Use the command `Combinations()` as follows:

```
> Combinations(5,3)
  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
    1    1    1    2    1    1    2    1    2    3
N    2    2    3    3    2    3    3    4    4    4
N    3    4    4    4    5    5    5    5    5    5
```

The 10 possible combinations are (1, 2, 3), (1, 2, 4), . . . , (3, 4, 5), listed vertically in the output. ■

Example 6.1 on the facing page assumed all of the values in the population of interest are sequential starting with the number one. It is not unusual to have non-sequential values for the population where the user desires to enumerate all possible combinations when sampling without replacement. To that end, code is provided (`SRS()`) that works in conjunction with `Combinations()` to list all of the possible combinations when using simple random sampling from a finite population:

```
> SRS <- function(POPvalues, n)
  { # SRS generates all possible SRS's of size n
    # from the population in vector POPvalues
    # by calling the function Combinations.
    N <- length(POPvalues)
    store <- t(Combinations(N, n))
    matrix(POPvalues[t(store)], nrow = nrow(store), byrow = TRUE) }
```

Example 6.2 Given a population of size $N = 5$, where $X_1 = 2$, $X_2 = 5$, $X_3 = 8$, $X_4 = 12$, and $X_5 = 13$, use `S` to list all of the possible samples of size $n = 3$. That is, list the $\binom{5}{3} = 10$ possible combinations.

Solution: First, make sure both the functions `Combinations()` and `SRS()` are stored on your computer by loading the `PASWR` package. Then, use the command `SRS()` as follows:

```
> t(SRS(c(2,5,8,12,13), 3))
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    2    2    2    5    2    2    5    2    5    8
[2,]    5    5    8    8    5    8    8   12   12   12
[3,]    8   12   12   12   13   13   13   13   13   13
```

The 10 possible combinations are (2, 5, 8), (2, 5, 12), . . . , (8, 12, 13), listed vertically in the output. The `S` command `t()` was used to transpose the data to conserve space. It is not obligatory to transpose the output; it is just as valid to type `SRS(c(2,5,8,12,13), 3)` so that the samples are listed across the rows instead of down the columns. ■

Example 6.3 A teacher wants an algorithm that will randomly select 5 students from a large lecture section of 180 students to present their work at the board.

Solution: Assume the students in the class are numbered from 1 to 180 according to the class roll and that the students know their numbers. Then, an unbiased procedure for selecting 5 students starts with using the following `S` code to determine which students should be in the sample:

```
> sample(1:180, 5, replace=FALSE)
[1] 138 52 135 58 160
```

Example 6.4 Randomly select 5 people from a group of 20 where the individuals are labeled from 1 to 20 and the individuals labeled 19 and 20 are four times more likely to be selected than the individuals labeled 1 through 18.

Solution: An unbiased procedure to select 5 people starts with using the following `S` code to determine which people will be in the sample:

```
> sample(x=(1:20), size=5, prob=c(rep(1/26,18), rep(4/26,2)), replace=FALSE)
[1] 20 19 1 17 16
```

6.1.2 Stratified Sampling

Simple random sampling gives samples that closely follow the population of interest provided the individual elements of the population of interest are relatively homogeneous with respect to the characteristics of interest in the study. When the population of interest is not homogeneous with respect to the characteristics under study, a possible solution might be to use **stratified sampling**.

Stratified sampling is most commonly used when the population of interest can be easily partitioned into subpopulations or strata. The strata are chosen to divide the population into non-overlapping, homogeneous regions. Then, the researcher takes simple random samples from each region or group. When using stratified sampling, it is crucial to select strata that are as homogeneous as possible within strata and as heterogeneous as possible between strata. For example, when agricultural researchers study crop yields, they tend to classify regions as arid and watered. It stands to reason that crop yields within arid regions will be poor and quite different from the yields from watered regions. Additional examples where stratified sampling can be used include:

1. In a study of the eating habits of a certain species, geographical areas often form natural strata.
2. In a study of political affiliation, gender often forms natural strata.
3. The Internal Revenue Service (IRS) might audit tax returns based on the reported taxable income by creating three groups: returns with reported taxable income less than \$ 50,000; returns with reported income less than \$75,000 but more than \$ 50,000; and returns with reported taxable income of more than \$ 75,000.

In addition to taking random samples within the strata, stratified samples are typically proportional to the size of their strata or proportional to the variability of the strata.

Example 6.5 A botanist wants to study the characteristics of a common weed and its adaptation to various geographical regions on a remote island. The island has well-defined strata that can be classified as desert, forest, mountains, and swamp. If 5000 acres of the island are desert, 1000 acres are forest, 500 acres are mountains, and 3500 acres are swamp, and the botanist wants to sample 5% of the population using a stratified sampling scheme that is proportional to the strata, how many acres of each of the four regions will he have to sample?

Solution: Since the size of the island is 10,000 acres, the botanist will need to sample a total of $10000 \times 0.05 = 500$ acres. The breakdown of the 500 acres is as follows: $500 \times \frac{5000}{10000} = 250$ desert acres; $500 \times \frac{1000}{10000} = 50$ forest acres; $500 \times \frac{500}{10000} = 25$ mountain acres; and $500 \times \frac{3500}{10000} = 175$ swamp acres. ■

6.1.3 Systematic Sampling

Systematic sampling is used when the researcher is in possession of a list that contains all N members of a given population and desires to select every k^{th} value in the master list. This type of sampling is often used to reduce costs since one only needs to select the initial starting point at random. That is, after the starting point is selected, the remaining values to be sampled are automatically specified.

To obtain a systematic sample, choose a sample size n and let k be the closest integer to $\frac{N}{n}$. Next, find a random integer i between 1 and k to be the starting point for sampling. Then, the sample is composed of the units numbered $i, i + k, i + 2k, \dots, i + (n - 1)k$. For

example, suppose a systematic sample is desired where 1 in $k = 100$ members is chosen from a list containing 1000 members. That is, every 100th member of the list is to be sampled. To pick the initial starting point, select a number at random between 1 and 100. If the random number generated is 53, then the researcher simply samples the values numbered 53, 153, 253, \dots , 953 from the master list. The following S code generates the locations to be sampled using a 1 in 100 systematic sampling strategy:

```
> seq(sample(1:100,1), 1000, 100)
[1] 53 153 253 353 453 553 653 753 853 953
```

Example 6.6 Produce a list of locations to sample for a systematic sample if $N = 1000$ and $n = 20$.

Solution: To take a systematic sample, every $k = \frac{1000}{20} = 50^{\text{th}}$ item will be observed. To start the process, select a random number between 1 and 50 using a random number generator. The following S code can be used to select a 1 in 50 systematic sample when $N = 1000$ and $k = 50$:

```
> seq(sample(1:50,1), 1000, 50)
[1] 27 77 127 ... 977
```

6.1.4 Cluster Sampling

Cluster sampling does not require a list of all of the units in the population like systematic sampling does. Rather, it takes units and groups them together to form clusters of several units. In contrast to stratified sampling, clusters should be as heterogeneous as possible within clusters and as homogeneous as possible between clusters. The main difference between cluster sampling and stratified sampling is that in cluster sampling, the cluster is treated as the sampling unit and analysis is done on a population of clusters. In one-step cluster sampling, all elements are selected in the chosen clusters. In stratified sampling, the analysis is done on elements within each strata. The main objective of cluster sampling is to reduce costs by increasing sampling efficiency. Examples of cluster sampling include:

1. Houses on a block
2. Students in school
3. Farmers in counties

6.2 Parameters

Once a sample is taken, the primary objective becomes to extract the maximum and most precise information as possible about the population from the sample. Specifically, the researcher is interested in learning as much as possible about the population's **parameters**. A parameter, θ , is a function of the probability distribution, F . That is, $\theta = t(F)$, where $t(\cdot)$ denotes the function applied to F . Each θ is obtained by applying some numerical procedure $t(\cdot)$ to the probability distribution function F . Although F has been used to denote the **cdf** exclusively until now, a more general definition of F is any description of \mathbf{X} 's probabilities. Note that the **cdf**, $\mathbb{P}(X \leq x)$, is included in this more general definition.

Parameters are what characterize probability distributions. More to the point, parameters are inherent in all probability models, and it is impossible to compute a probability without prior knowledge of the distribution's parameters. Parameters are treated as constants in classical statistics and as random variables in Bayesian statistics. In everything that follows, parameters are treated as constants.

Example 6.7 Suppose F is the exponential distribution, $F = \text{Exp}(\lambda)$, and $t(F) = E_F(\mathbf{X}) = \theta$. Express θ in terms of λ .

Solution: Here, $t(\cdot)$ is the expected value of \mathbf{X} , so $\theta = 1/\lambda$. ■

6.2.1 Infinite Populations' Parameters

The most commonly estimated parameters are the mean (μ), the variance (σ^2), and the proportion (π). What follows is a brief review of their definitions.

Population mean — The **mean** is defined as the expected value of the random variable X .

- If X is a discrete random variable,

$$\mu_X = E[X] = \sum_{i=1}^{\infty} x_i \mathbb{P}(X = x_i), \text{ where } \mathbb{P}(X = x_i) \text{ is the pdf of } X.$$

- If X is a continuous random variable,

$$\mu_X = E[X] = \int_{-\infty}^{\infty} x f(x) dx, \text{ where } f(x) \text{ is the pdf of } X.$$

Population variance — The population variance is defined as $\text{Var}[X] = E[(X - \mu)^2]$.

- For the discrete case,

$$\sigma_X^2 = \text{Var}[X] = \sum_{i=1}^{\infty} (x_i - \mu)^2 \cdot \mathbb{P}(X = x_i) = \sum_{i=1}^{\infty} x_i^2 \cdot \mathbb{P}(X = x_i) - \mu^2.$$

- For the continuous case,

$$\sigma_X^2 = \text{Var}[X] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2.$$

Population proportion — The population proportion π is the ratio

$$\pi = \frac{N_1}{N},$$

where N_1 is the number of values that fulfill a particular condition and N is the size of the population.

6.2.2 Finite Populations' Parameters

Suppose a finite population that consists of N elements, X_1, \dots, X_N , is defined. The most commonly defined parameters are in Table 6.1 on the next page.

Table 6.1: Finite populations' parameters

Population Parameter	Formula	Explanation
Mean	$\mu_f = \frac{\sum_{i=1}^N X_i}{N}$	
Total	$\tau = \sum_{i=1}^N X_i = N\mu_f$	
Proportion	$\pi_f = \frac{Y}{N}$	Where Y is the number of elements of the population that fulfill a certain characteristic.
Proportion (alternate)	$\pi_f = \frac{\sum_i Y_i}{N}$	The Y_i s take on a value of 1 if they represent a certain characteristic and 0 if they do not possess the characteristic
Variance(N)	$\begin{aligned} \sigma_{f;N}^2 &= \frac{\sum_{i=1}^N (X_i - \mu_f)^2}{N} \\ &= \frac{1}{N} \sum_{i=1}^N X_i^2 - (\mu_f)^2 \end{aligned}$	
Variance ($N - 1$)	$\sigma_{f;N-1}^2 = \frac{\sum_{i=1}^N (X_i - \mu_f)^2}{N - 1}$	
Variance (dichotomous)	$\sigma_f^2 = \pi_f(1 - \pi_f)$	π_f represents the proportion of elements in the population with a common characteristic
Standard Deviation	$\sigma_f = \sqrt{\frac{\sum_{i=1}^N (X_i - \mu_f)^2}{N}}$	

6.3 Estimators

Population parameters are generally unknown. Consequently, one of the first tasks is to estimate the unknown parameters using sample data. Estimates of the unknown parameters are computed with **estimators** or **statistics**. An estimator is a function of the sample, while an estimate (a number) is the realized value of an estimator that is obtained when a sample is actually taken. Given a random sample, $\{X_1, X_2, \dots, X_n\} = \mathbf{X}$, from a probability distribution F , a statistic, any function of the sample is denoted as $T = t(\mathbf{X})$. Note that the estimator T of θ will at times also be denoted $\hat{\theta}$. Since a statistic is a function of the random variables \mathbf{X} , it follows that statistics are also random variables. The specific value of a statistic can only be known after a sample has been taken. The resulting number, computed from a statistic, is called an **estimate**. For example, the arithmetic mean of a sample

$$T = t(\mathbf{X}) = \bar{X} = \frac{\sum_{i=1}^n X_i}{n}, \quad (6.1)$$

is a statistic (estimator) constructed from a random sample $\{X_1, \dots, X_n\}$.

Until a sample is taken, the value of the statistic (the estimate) is unknown. Suppose a random sample has been taken that contains the following values: $\mathbf{x} = \{3, 5, 6, 1, 2, 7\}$. It follows that the value of the statistic $T = t(\mathbf{X})$, where $t(\mathbf{X})$ is defined in (6.1) as $t = t(\mathbf{x}) = \frac{3+5+6+1+2+7}{6} = 4$. The quantity $t(\mathbf{X}) = \frac{X_1 \times X_2}{6}$ is also a statistic; however, it does not have the same properties as the arithmetic mean defined in (6.1).

The essential distinction between parameters and estimators is that a parameter is a constant in classical statistics while an estimator is a random variable, since its value changes from sample to sample. Parameters are typically designated with lowercase Greek letters, while estimators are typically denoted with lowercase Latin letters. However, when working with finite populations, it is standard notation to use different capital Latin letters to denote both parameters and estimators. At times, it is also common to denote an estimator by placing a hat over a parameter such as $\hat{\beta}_1$. Some common parameters and their corresponding estimators are provided in Table 6.2.

Table 6.2: Parameters and their corresponding estimators

Parameter	Name	Estimator (Latin notation)	Estimator (Hat notation)
μ	population mean	\bar{X} sample mean	$\hat{\mu}$
σ^2	population variance	S^2 sample variance	$\hat{\sigma}^2$

Some of the statistics used to estimate parameters when sampling from a finite population are given in Table 6.3 while the more common statistics used when working with a random sample of size n are given in Table 6.4 on the facing page.

Table 6.3: Finite population parameter estimators and their standard errors

Parameter	Estimator	$\hat{\sigma}_{\text{estimator}}$
Population Mean	$\bar{X}_f = \frac{\sum_{i=1}^n X_i}{n}$	$\frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N}}$
Population Total	$T_f = N \bar{X}_f$	$\frac{S}{\sqrt{n}} \cdot N \cdot \sqrt{\frac{N-n}{N}}$
Population Proportion	$P = \frac{Y}{n}$	$\sqrt{\frac{P(1-P)}{n-1} \left(\frac{N-n}{N} \right)}$

6.3.1 Empirical Probability Distribution Function

The **empirical probability distribution function**, **epdf** $= \hat{F}$, is defined as the discrete distribution that puts probability $\frac{1}{n}$ on each value in \mathbf{x} , where \mathbf{x} is a sample of size n extracted from F . The **empirical cumulative distribution function**, **ecdf**, is defined as

$$\hat{F}_n(t) = \sum_{i=1}^n \mathbf{I}\{x_i \leq t\} / n. \quad (6.2)$$

Here, $\mathbf{I}\{x_i \leq t\}$ is the indicator function that returns a value of 1 when $x_i \leq t$ and 0 when $x_i > t$.

Table 6.4: Statistics for samples of size n

Statistic	Formula	Explanation
Mean	$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$	
Total	$T = n\bar{X}$	
Proportion	$P = \frac{Y}{n}$	Where Y is the number of elements with a certain characteristic
Variance (uncorrected)	$S_u^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$ $= \frac{\sum_{i=1}^n X_i^2}{n} - \bar{X}^2$	
Variance	$S_{ud}^2 = P(1 - P)$	Uncorrected and dichotomous
Variance	$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$ $= \frac{n}{n - 1} S_u^2$	
Variance (dichotomous)	$S_d^2 = \frac{nP(1 - P)}{n - 1}$	If $n \geq 20$, S^2 can be approximated with the quantity $P(1 - P)$.
Standard Deviation	$S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}}$	

Example 6.8 Simulate rolling a die 100 times and compute the **epdf**. Graph the **ecdf**.

Solution: The R code to solve the problem is

```
> rolls <- sample(1:6,100, replace=TRUE)
> table(rolls)
rolls
 1  2  3  4  5  6
22 18 12 16 15 17

> table(rolls)/100      # epdf
rolls
 1  2  3  4  5  6
0.22 0.18 0.12 0.16 0.15 0.17
> plot(ecdf(rolls))
```

where the output following `table(rolls)/100` is the empirical distribution function. The graph of the realized **ecdf** is found in Figure 6.1 on the next page.

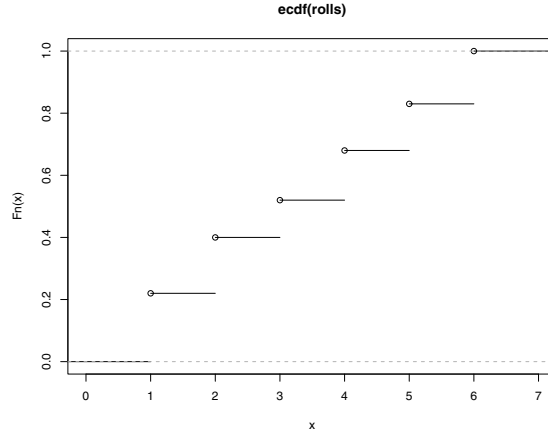


FIGURE 6.1: Empirical cumulative distribution function of rolling a die 100 times ■

6.3.2 Plug-In Principle

The **plug-in principle** is an intuitive method of estimating parameters from samples. The **plug-in estimator** of a parameter $\theta = t(F)$ is defined to be $\hat{\theta} = t(\hat{F})$. Simply put, the estimate is the result of applying the function $t(\cdot)$ to the empirical probability distribution \hat{F} .

Example 6.9 What are the plug-in estimators of (a) the expected value and (b) the variance of a discrete distribution F ?

Solution: The answers are as follows:

- (a) When the expected value is $\theta = E_F(\mathbf{X})$, the plug-in estimator of the expected value is $\hat{\theta} = E_{\hat{F}}(\mathbf{X}) = \sum_{i=1}^n X_i \cdot \frac{1}{n} = \bar{X}$.
- (b) When the variance is $\theta = Var_F(\mathbf{X}) = E_F(\mathbf{X} - \mu)^2$, the plug-in estimator of the variance of \mathbf{X} is $\hat{\theta} = E_{\hat{F}}(X - \bar{X})^2 = \sum_{i=1}^n (X_i - \bar{X})^2 \cdot \frac{1}{n}$. ■

6.4 Sampling Distribution of \bar{X}

Suppose 10 college students are randomly selected from the population of college students in the state of Colorado and compute the mean age of the sampled students. If this process were repeated three times, it is unlikely any of the computed sample means would be identical. Likewise, it is not likely that any of the three computed sample means would be exactly equal to the population mean. However, these sample means are typically used to estimate the unknown population mean. So, how can the accuracy of the sampled value be assessed?

To assess the accuracy of a value (estimate) returned from a statistic, the probability distribution of the statistic of interest is used to place probabilistic bounds on the sampling error. The probability distribution associated with all of the possible values a statistic can assume is called the **sampling distribution** of the statistic. This section presents the sampling distribution of the sample mean. Before discussing the sampling distribution of \bar{X} , the mean and variance of \bar{X} for any random variable X are highlighted.

If X is a random variable with mean μ and variance σ^2 , and if a random sample X_1, \dots, X_n is taken, the expected value and variance of \bar{X} are written

$$E[\bar{X}] = \mu_{\bar{X}} = \mu, \quad (6.3)$$

$$\text{Var}[\bar{X}] = \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}. \quad (6.4)$$

The computations of the answers for (6.3) and (6.4) are the same as those for Example 5.9 on page 180, which are reproduced for the reader's benefit:

$$E[\bar{X}] = \sum_{i=1}^n \frac{E[X_i]}{n} = \sum_{i=1}^n \frac{1}{n} \mu = \mu,$$

$$\text{Var}[\bar{X}] = \text{Var}\left[\sum_{i=1}^n \frac{X_i}{n}\right] = \frac{1}{n^2} \sum_{i=1}^n \text{Var}[X_i] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Clearly, as the sample size increases, the variance of the sampling distribution of \bar{X} decreases.

Example 6.10 ▷ *Sampling: Balls in an Urn* ◁ Consider an experiment where two balls are randomly selected from an urn containing six numbered balls. First, the sampling is done with replacement (Case 1), and then the sampling is done without replacement (Case 2). List the exact sampling distributions of \bar{X} and S^2 for both cases. Finally, create graphs that compare these four distributions.

Solution: **Case 1** When the sampling is performed with replacement, the outcomes can be viewed as a random sample of size 2 drawn from a discrete uniform distribution. The mean and variance of the uniform distribution are

$$\mu = \frac{1 + 2 + \dots + 6}{6} = 3.5$$

and

$$\sigma^2 = E(X^2) - \mu^2 = \frac{1^2 + 2^2 + \dots + 6^2}{6} - (3.5)^2 = 2.9166.$$

Note that these values could also be computed using the formulas $\mu = (N + 1)/2$ and $\sigma^2 = (N^2 - 1)/12$ given in (4.9).

There are 36 possible samples of size 2 from this distribution listed in Table 6.5 on the next page. Using the fact that each of samples listed in Table 6.5 is equally likely (1/36), construct both the sampling distribution of \bar{X} given in Table 6.6 on the following page and the sampling distribution of S^2 given in Table 6.7 on the next page.

The mean of the sampling distribution, $\mu_{\bar{X}} = E[\bar{X}]$, and the variance of the sampling distribution, $\sigma_{\bar{X}}^2 = E[(\bar{X} - \mu_{\bar{X}})^2]$, are

$$\mu_{\bar{X}} = E[\bar{X}] = 1 \times \frac{1}{36} + 1.5 \times \frac{2}{36} + \dots + 6 \times \frac{1}{36} = 3.5$$

and

$$\begin{aligned} \sigma_{\bar{X}}^2 = E[(\bar{X} - \mu_{\bar{X}})^2] &= (1 - 3.5)^2 \times \frac{1}{36} + (1.5 - 3.5)^2 \times \frac{2}{36} + \\ &\quad \dots + (6 - 3.5)^2 \times \frac{1}{36} = 1.4583. \end{aligned}$$

Table 6.5: Possible samples of size 2 with \bar{x} and s^2 for each sample – random sampling

(x_1, x_2)	\bar{x}	s^2	(x_1, x_2)	\bar{x}	s^2
(1, 1)	1.0	0.0	(4, 1)	2.5	4.5
(1, 2)	1.5	0.5	(4, 2)	3.0	2.0
(1, 3)	2.0	2.0	(4, 3)	3.5	0.5
(1, 4)	2.5	4.5	(4, 4)	4.0	0.0
(1, 5)	3.0	8.0	(4, 5)	4.5	0.5
(1, 6)	3.5	12.5	(4, 6)	5.0	2.0
(2, 1)	1.5	0.5	(5, 1)	3.0	8.0
(2, 2)	2.0	0.0	(5, 2)	3.5	4.5
(2, 3)	2.5	0.5	(5, 3)	4.0	2.0
(2, 4)	3.0	2.0	(5, 4)	4.5	0.5
(2, 5)	3.5	4.5	(5, 5)	5.0	0.0
(2, 6)	4.0	8.0	(5, 6)	5.5	0.5
(3, 1)	2.0	2.0	(6, 1)	3.5	12.5
(3, 2)	2.5	0.5	(6, 2)	4.0	8.0
(3, 3)	3.0	0.0	(6, 3)	4.5	4.5
(3, 4)	3.5	0.5	(6, 4)	5.0	2.0
(3, 5)	4.0	2.0	(6, 5)	5.5	0.5
(3, 6)	4.5	4.5	(6, 6)	6.0	0.0

Table 6.6: Sampling distribution of \bar{X} – random sampling

\bar{x}	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6
$f(\bar{x})$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Table 6.7: Sampling distribution of S^2 – random sampling

s^2	0	0.5	2	4.5	8	12.5
$f(s^2)$	6/36	10/36	8/36	6/36	4/36	2/36

Note that the computed values of $E[\bar{X}]$ and $\sigma_{\bar{X}}^2$ are in agreement with the formulas $E[\bar{X}] = \mu$ and $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$ given in (6.3) and (6.4). Also note that $E[S^2] = \sigma^2$. Specifically,

$$E[S^2] = 0 \times \frac{6}{36} + 0.5 \times \frac{10}{36} + \dots + 12.5 \times \frac{2}{36} = 2.9166.$$

Case 2 When the sampling is performed without replacement, the outcomes can be viewed as a **simple random sample** of size 2 drawn from a discrete uniform distribution. Note

that fewer samples exist when sampling without replacement ($\binom{6}{2} = 15$), but that each sample is equally likely to be drawn. The 15 possible samples of size 2 from this distribution are listed in Table 6.8. Using the fact that each of the samples listed in Table 6.8 is equally likely ($1/15$), construct the sampling distribution of \bar{X} given in Table 6.9, and the sampling distribution of S^2 given in Table 6.10 both on the current page.

Table 6.8: Possible samples of size 2 with \bar{x} and s^2 – simple random sampling

(x_1, x_2)	\bar{x}	s^2
(1, 2)	1.5	0.5
(1, 3)	2	2.0
(1, 4)	2.5	4.5
(1, 5)	3	8.0
(1, 6)	3.5	12.5
(2, 3)	2.5	0.5
(2, 4)	3	2.0
(2, 5)	3.5	4.5
(2, 6)	4	8.0
(3, 4)	3.5	0.5
(3, 5)	4	2.0
(3, 6)	4.5	4.5
(4, 5)	4.5	0.5
(4, 6)	5	2.0
(5, 6)	5.5	0.5

Table 6.9: Sampling distribution of \bar{X} – simple random sampling

\bar{x}	1.5	2	2.5	3	3.5	4	4.5	5	5.5
$f(\bar{x})$	1/15	1/15	2/15	2/15	3/15	2/15	2/15	1/15	1/15

Table 6.10: Sampling distribution of S^2 – simple random sampling

s^2	0.5	2	4.5	8	12.5
$f(s^2)$	5/15	4/15	3/15	2/15	1/15

The mean of the sampling distribution, $\mu_{\bar{X}} = E[\bar{X}]$, the variance of the sampling distribution, $\sigma_{\bar{X}}^2 = E[(\bar{X} - \mu_{\bar{X}})^2]$, and the expected value of S^2 , $E[S^2]$, are

$$\begin{aligned}\mu_{\bar{X}} &= E[\bar{X}] = 1.5 \times \frac{1}{15} + 2 \times \frac{1}{15} + \cdots + 5.5 \times \frac{1}{15} = 3.5, \\ \sigma_{\bar{X}}^2 &= E[(\bar{X} - \mu_{\bar{X}})^2] = (1.5 - 3.5)^2 \times \frac{1}{15} + (2 - 3.5)^2 \times \frac{1}{15} + \\ &\quad \cdots + (5.5 - 3.5)^2 \times \frac{1}{15} = 1.16666, \\ \text{and } E[S^2] &= 0.5 \times \frac{5}{15} + 2 \times \frac{4}{15} + \cdots + 12.5 \times \frac{1}{15} = 3.5.\end{aligned}$$

Remarkably, the sample mean is identical when sampling with and without replacement. In fact, the expected value of the sample mean is μ whether sampling with or without replacement. The variance of the sample mean and the expected value of the sample variance have changed, however. These changes are due to the fact that sampling is from a finite population without replacement. A summary of the formulas used to compute these results is found in Table 6.11.

Table 6.11: Summary results for sampling without replacement (finite population)

$\mu_{\bar{X}} = \mu_f$
$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$
$E[S^2] = \frac{N}{N-1} \cdot \sigma^2$
$E[S_u^2] = \frac{N}{N-1} \cdot \frac{n-1}{n} \cdot \sigma^2$

Note that the computed values of $E[\bar{X}] = \mu_f = 3.5$, $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1} = \frac{2.9166}{2} \cdot \frac{6-2}{6-1} = 1.1666$, and $E[S^2] = \frac{N}{N-1} \sigma^2 = \frac{6}{5}(2.9166) = 3.5$ for this example are in agreement with the formulas for sampling without replacement given in Table 6.11. A comparison of the results from Case 1 and Case 2 can be found in Table 6.12.

Table 6.12: Computed values for random sampling (Case 1) and simple random sampling (Case 2)

	μ	$E[\bar{X}]$	σ^2	$E[S^2]$	$\sigma_{\bar{X}}^2$
Case 1	3.5	3.5	2.9166	2.9166	1.4583
Case 2	3.5	3.5	2.9166	3.5	1.1666

Graphical comparisons for the sampling distributions of \bar{X} and S^2 when sampling with replacement (random sampling) and when sampling without replacement (simple random sampling) are depicted in Figure 6.2 on the next page. The following S code can be used to verify all the results in this solution:

```
> N <- 6
> n <- 2
> pop <- 1:N
> rs <- expand.grid(Draw1=pop, Draw2=pop) # Possible random samples
> xbarN <- apply(rs, 1, mean) # Means of all rs values
> s2N <- apply(rs, 1, var) # Variance of all rs values
> TOT1 <- cbind(rs, xbarN=xbarN, s2N=s2N)

> TOT1 # Numerical values for Table 6.5

> table(xbarN) # Numerators for Table 6.6

> table(s2N) # Numerators for Table 6.7

> MU <- mean(pop) # Population mean
> VAR <- sum((pop-mean(pop))^2)*(1/N) # Population variance
> MU.xbarN <- mean(xbarN) # Expected value of xbarN
> E.s2N <- mean(s2N) # Expected value of s2N
> VAR.xbarN <- sum((xbarN-mean(xbarN))^2)*(1/(N*N))
> reN <- c(MU, MU.xbarN, VAR, E.s2N, VAR.xbarN)
> names(reN) <- c("MU", "MU.xbarN", "VAR", "E.s2N", "V.xbarN")

> reN # Numerical values for Case 1 in Table 6.12

> srs <- SRS(1:N, n) # Possible simple random samples
> xbari <- apply(srs, 1, mean) # Means of simple random samples
> s2i <- apply(srs, 1, var) # Variances of simple random samples
> TOT <- cbind(srs, xbari, s2i)
> dimnames(TOT)[[2]] <- c("Draw1", "Draw2", "xbari", "s2i")

> TOT # Numerical values for Table 6.8

> table(xbari) # Numerators for Table 6.9

> table(s2i) # Numerators for Table 6.10

> MU <- mean(pop) # Population mean
> VAR <- sum((pop-mean(pop))^2)*(1/N) # Population variance
> MU.xbar <- mean(xbari) # Expected value of xbari
> E.s2 <- mean(s2i) # Expected value of s2i
> VAR.xbar <- sum((xbari-mean(xbari))^2)*(1/choose(N, n))
> results <- c(MU, MU.xbar, VAR, E.s2, VAR.xbar)
> names(results) <- c("MU", "MU.xbari", "VAR", "E.s2i", "V.xbari")

> print(results) # Numerical values for Case 2 in Table 6.12
```

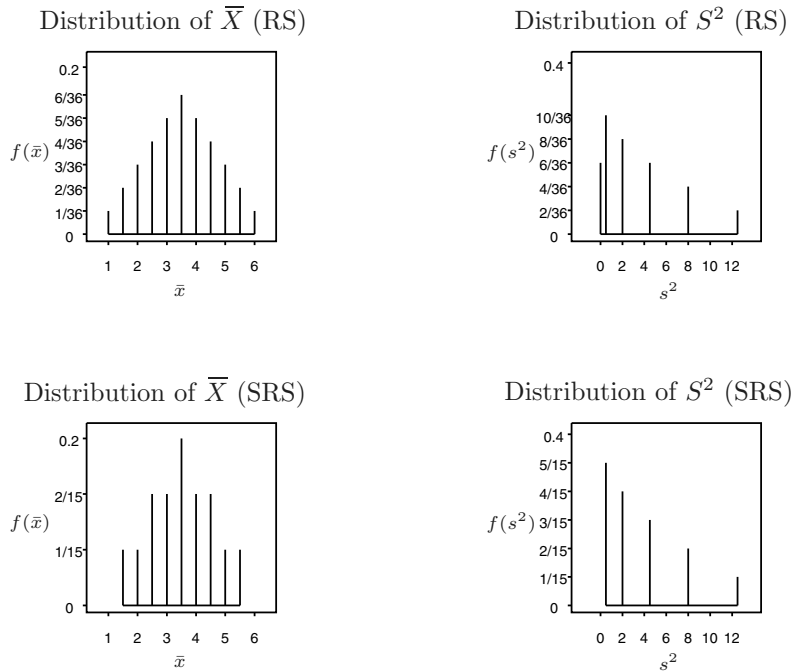


FIGURE 6.2: Sampling distributions of \bar{X} and S^2 under random sampling (RS) and simple random sampling (SRS) for Example 6.10 on page 207 are given. Note that the dispersion for the sampling distribution of \bar{X} is smaller under Case 2 than it is with Case 1. ■

6.5 Sampling Distribution for a Statistic from an Infinite Population

Consider a population from which k random samples, each of size n , are taken. In general, if given k samples, k different values for the sample mean will result. If k is very large, theoretically infinite, the values of the means from each of the samples, denoted \bar{X}_i for each sample i , will be random variables with a resulting distribution referred to as the sampling distribution of the sample mean. The sampling distribution of a statistic, $t(X)$, is the resulting probability distribution for $t(X)$ calculated by taking an infinite number of random samples of size n . The resulting sampling distribution will typically not coincide with the distribution of the parent population.

6.5.1 Sampling Distribution for the Sample Mean

6.5.1.1 First Case: Sampling Distribution of \bar{X} when Sampling from a Normal Distribution

When sampling from a normal distribution, the resulting sampling distribution for the sample mean is also a normal distribution. This is an immediate result of Theorem 5.1 on page 176. That is, \bar{X} is a linear combination of the X_i s where $a_i = \frac{1}{n}$. As observed earlier, the mean and the variance of the sampling distribution of \bar{X} are μ and σ^2/n regardless of

the underlying population. So, the mean and variance of the sampling distribution of \bar{X} are always known. However, it is not always true that the resulting sampling distribution of \bar{X} is known. If $X \sim N(\mu, \sigma)$, then $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$.

Example 6.11 If $X \sim N(\mu, 12)$, find the required sample size to guarantee $|\bar{X} - \mu| < 3$ with a probability of 0.95.

Solution: Changing the prose into a mathematical statement,

$$\mathbb{P}(|\bar{X} - \mu| < 3) = 0.95$$

needs to be solved.

Since $X \sim N(\mu, \sigma = 12)$, it follows that

$$\bar{X} \sim N\left(\mu, \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{n}}\right).$$

Consequently,

$$\mathbb{P}\left(\frac{|\bar{X} - \mu|}{\sigma/\sqrt{n}} < 1.96\right) = 0.95.$$

Multiplying both sides by $\frac{\sigma}{\sqrt{n}}$ and substituting 12 for σ gives

$$\mathbb{P}\left(|\bar{X} - \mu| < (1.96)\frac{12}{\sqrt{n}}\right) = 0.95.$$

Multiplying both sides by \sqrt{n} , dividing both sides by 3, and finally squaring both sides, gives $n = 61.47$. Consequently, a sample size of at least 62 is needed to guarantee $|\bar{X} - \mu| < 3$ with a probability of 0.95. ■

Example 6.12 A small town in the Pyrenean mountains wants to reduce the bear population because several sheep have recently been killed by bears. Three autonomous communities (Cataluña, Aragón, and Navarra) have made bids to remove 10 bears. The three autonomous communities indicated in their bids that they are willing to spend 5, 7.5, and 10 thousand dollars per bear to capture the bears. Decide which autonomous communities can capture 10 bears with a probability of at least 0.999 knowing that the cost to capture a bear follows a normal distribution with a mean of 5 thousand dollars and a standard deviation of 0.6 thousand dollars.

Solution: Assuming that the costs to capture the bears act as independent random variables, such that if X_i is the cost to capture one bear, the total cost to capture 10 bears is also a random variable, given by $Y = X_1 + \cdots + X_{10}$. Since $X_i \sim N(5, 0.6)$, it follows using Theorem 5.1 on page 176 that the mean of Y will be $5 \cdot 10 = 50$ and the standard deviation of Y will be $\sqrt{10 \cdot (0.6)^2} = 1.897367$. Mathematically, write $Y \sim N(50, 1.897367)$. Cataluña will be able to capture 10 bears provided $Y \leq 50$, Aragón will be able to capture 10 bears provided $Y \leq 75$, and Navarra will be able to capture 10 bears provided $Y \leq 100$. The probabilities of these events are

$$\mathbb{P}(Y \leq 50) = \mathbb{P}\left(Z \leq \frac{50 - 50}{1.897367}\right) = \mathbb{P}(Z \leq 0) = 0.5,$$

$$\mathbb{P}(Y \leq 75) = \mathbb{P}\left(Z \leq \frac{75 - 50}{1.897367}\right) = \mathbb{P}(Z \leq 13.17616) = 1,$$

$$\text{and } \mathbb{P}(Y \leq 100) = \mathbb{P}\left(Z \leq \frac{100 - 50}{1.897367}\right) = \mathbb{P}(Z \leq 26.35231) = 1.$$

The following S code computes the answers directly:

```
> pnorm(50,50,1.897367)
[1] 0.5
> pnorm(75,50,1.897367)
[1] 1
> pnorm(100,50,1.897367)
[1] 1
```

There is only a 50% chance that the Catalan bid would provide sufficient funds to catch 10 bears. On the other hand, the bids from Navarra and Aragón would both have a 100% chance of catching all 10 bears. ■

Example 6.13 It is well-known that the measurement errors committed by employees when they measure the length of a zipper in a particular assembly process follow a normal distribution with a mean of 0 and standard deviation of 2 millimeters. Find

- The maximum error for measuring a zipper a single time with 0.95 probability.
- The maximum error of the mean measurement of the zipper with 0.95 probability if it is measured 10 times.
- The number of times one needs to measure a zipper to ensure the maximum measurement error of the mean is less than 1 millimeter with 0.95 probability.

Solution: The solutions are as follows:

(a) Let the random variable X represent the measurement error committed by employees when measuring zippers. Since $X \sim N(0, 2)$, $Z = \frac{X}{2} \sim N(0, 1)$. Since

$$\mathbb{P}(-1.96 < Z < 1.96) = 0.95,$$

and since $Z = \frac{X}{2}$,

$$\mathbb{P}(-1.96 < \frac{X}{2} < 1.96) = 0.95.$$

Basic algebra then gives

$$|X| < 2(1.96) = 3.92.$$

(b) In this question, the distribution of X is no longer the focus, but rather the distribution of \bar{X} is. Since $\bar{X} \sim N(0, 2/\sqrt{10})$, it follows that the maximum error committed when measuring a zipper 10 times is

$$|\bar{X}| = \frac{2}{\sqrt{10}}(1.96) = 1.24.$$

(c) Since $\bar{X} \sim N(0, 2/\sqrt{n})$, it follows that

$$|\bar{X}| = \frac{2}{\sqrt{n}}(1.96) \leq 1$$

must be solved for n . The solution is $n \geq (3.92)^2 = 15.36$. In other words, at least 16 zippers must be measured to ensure the maximum measurement error of the mean is no more than 1 millimeter with 0.95 probability. ■

6.5.1.2 Second Case: Sampling Distribution of \bar{X} when X Is not a Normal Random Variable

When the underlying population of X is not normal, provided the sample size is sufficiently large, the sampling distribution of \bar{X} is still normal. Specifically, the **Central Limit Theorem** states that if $X \sim (\mu, \sigma)$, then the limiting distribution of

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

as $n \rightarrow \infty$ is the standard normal distribution. Expressed in lay terms, the sampling distribution of \bar{X} , regardless of the underlying population, is approximately $N(\mu, \sigma/\sqrt{n})$ provided n is sufficiently large. Populations that are asymmetric require larger values of n compared to symmetric populations before the sampling distribution of \bar{X} appears normal.

Consider the left graph of Figure 6.3, which depicts a $Unif(0, 10)$ population, while the center graph of Figure 6.3 depicts the theoretical sampling distribution of \bar{X} for samples of size $n = 2$ when sampling is from a $Unif(0, 10)$ population. Finally, the far right graph of Figure 6.3 superimposes the theoretical sampling distribution of \bar{X} for samples of size $n = 2$ when sampling is from a $Unif(0, 10)$ population over a normal distribution with a mean and standard deviation corresponding to the mean and standard deviation of the sampling distribution of \bar{X} for samples of size $n = 2$ when sampling from the $Unif(0, 10)$ population. It is interesting to note in the far right graph in Figure 6.3, how closely the triangular distribution resembles the normal distribution.

The sampling distributions of \bar{X} associated with infinite populations are obviously impossible to enumerate. However, simulation can be used to gain insight into the sampling distribution of \bar{X} when sampling from known populations. That is, a large number of samples from a known population can be drawn and the distribution of \bar{X} can be studied.

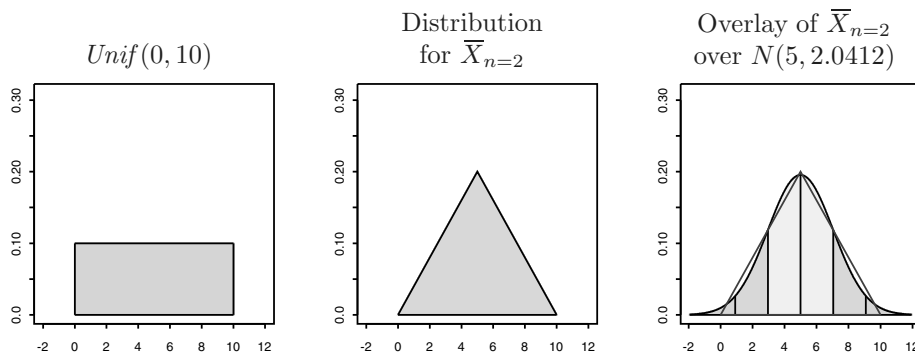


FIGURE 6.3: The far left graph depicts a $Unif(0, 10)$ distribution. The middle graph depicts the theoretical sampling distribution of \bar{X} for samples of size $n = 2$ when the samples are drawn from a $Unif(0, 10)$ distribution. The far left graph depicts a $N(5, 2.0412)$ distribution overlaid with the theoretical distribution of \bar{X} for samples of size $n = 2$ when the samples are drawn from a $Unif(0, 10)$ distribution.

In what follows, the various graphs depicted in Figure 6.4 on the next page and Figure 6.5 on page 217 are examined to gain insight into how large the sample size, n , needs to be when working with both symmetric distributions and skewed distributions such as the uniform

distribution and the exponential distribution, respectively. S is used to simulate $m = 50,000$ samples of sizes $n = 2, 16, 36,$ and 100 from a $Unif(-3.66025, 13.66025)$ distribution and an $Exp(5)$ distribution. Note that both the means and standard deviations are 5 and 5 for these distributions.

Figure 6.4 depicts the simulated sampling distribution of \bar{X} for samples of sizes $n = 2$ and 16 when one samples from a $Unif(-3.66025, 13.66025)$ distribution and an $Exp(5)$ distribution, respectively. Figure 6.5 on the facing page depicts the simulated sampling distribution of \bar{X} for samples of sizes $n = 36$ and 100 when sampling from a $Unif(-3.66025, 13.66025)$ distribution and an $Exp(5)$ distribution, respectively. What should become evident from looking at Figures 6.4 and 6.5 is that the sampling distribution of \bar{X} when sampling from a uniform distribution becomes approximately normal much sooner than does the sampling distribution of \bar{X} when sampling from an exponential distribution.

In addition to assessing the simulated sampling distributions of \bar{X} graphically by superimposing a normal density with mean and standard deviation equal to the mean and standard deviation of the sampling distribution of \bar{X} as shown in Figures 6.4 and 6.5, Table 6.13 on the next page is provided which contains the percent of the simulated sampling distribution of \bar{X} that falls within $(-\infty, \mu_{\bar{X}} - 2\sigma_{\bar{X}}]$, $(\mu_{\bar{X}} - 2\sigma_{\bar{X}}, \mu_{\bar{X}} - \sigma_{\bar{X}}]$, $(\mu_{\bar{X}} - \sigma_{\bar{X}}, \mu_{\bar{X}}]$, $(\mu_{\bar{X}}, \mu_{\bar{X}} + \sigma_{\bar{X}}]$, $(\mu_{\bar{X}} + \sigma_{\bar{X}}, \mu_{\bar{X}} + 2\sigma_{\bar{X}}]$, and $(\mu_{\bar{X}} + 2\sigma_{\bar{X}}, \infty]$ for sample sizes $n = 2, 16, 36,$ and 100 when sampling from a $Unif(-3.66025, 13.66025)$ distribution and an $Exp(5)$ distribution. By studying the percentages from the simulations in Table 6.13 on the facing page, one can see that the simulated sampling distribution of \bar{X} when sampling from an exponential distribution is still slightly skewed even for sample sizes as large as $n = 100$. To verify the numbers presented in Table 6.13 on the next page and to create graphs similar to those in Figures 6.4 and 6.5, the user can use the code `n2UNIFsim` provided at <http://www1.appstate.edu/~arnholta/PASWR> in the Chapter 6 script.

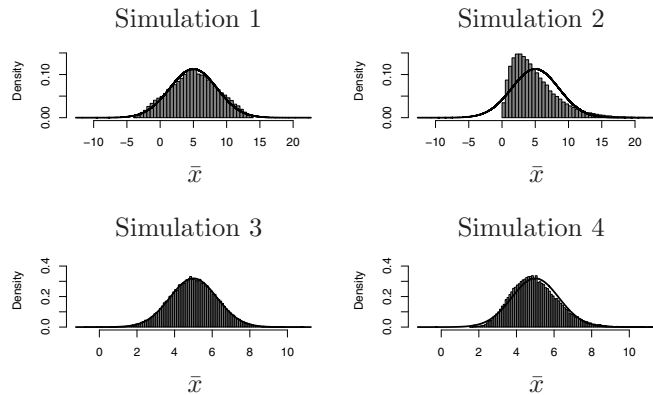


FIGURE 6.4: Simulation 1 depicts the simulated sampling distribution of \bar{X} for samples of size $n = 2$ that are selected from a $Unif(-3.66025, 13.66025)$ distribution. Simulation 2 depicts the simulated sampling distribution of \bar{X} for samples of size $n = 2$ that are selected from an $Exp(5)$ distribution. Simulation 3 depicts the simulated sampling distribution of \bar{X} for samples of size $n = 16$ that are selected from a $Unif(-3.66025, 13.66025)$ distribution. Simulation 4 depicts the simulated sampling distribution of \bar{X} for samples of size $n = 16$ that are selected from an $Exp(5)$.

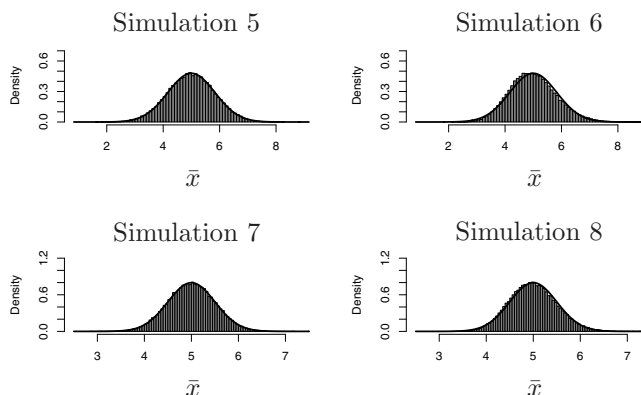


FIGURE 6.5: Simulation 5 depicts the simulated sampling distribution of \bar{X} for samples of size $n = 36$ that are selected from a $Unif(-3.66025, 13.66025)$ distribution. Simulation 6 depicts the simulated sampling distribution of \bar{X} for samples of size $n = 36$ that are selected from an $Exp(5)$ distribution. Simulation 7 depicts the simulated sampling distribution of \bar{X} for samples of size $n = 100$ that are selected from a $Unif(-3.66025, 13.66025)$ distribution. Simulation 8 depicts the simulated sampling distribution of \bar{X} for samples of size $n = 100$ that are selected from an $Exp(5)$.

Table 6.13: Comparison of simulated uniform and exponential distributions to the normal distribution, $Int1 = (-\infty, \mu_{\bar{X}} - 2\sigma_{\bar{X}}]$, $Int2 = (\mu_{\bar{X}} - 2\sigma_{\bar{X}}, \mu_{\bar{X}} - \sigma_{\bar{X}}]$, $Int3 = (\mu_{\bar{X}} - \sigma_{\bar{X}}, \mu_{\bar{X}}]$, $Int4 = (\mu_{\bar{X}}, \mu_{\bar{X}} + \sigma_{\bar{X}}]$, $Int5 = (\mu_{\bar{X}} + \sigma_{\bar{X}}, \mu_{\bar{X}} + 2\sigma_{\bar{X}}]$, $Int6 = (\mu_{\bar{X}} + 2\sigma_{\bar{X}}, \infty]$

		$Int1$	$Int2$	$Int3$	$Int4$	$Int5$	$Int6$
	$N(0, 1)$	0.0228	0.1359	0.3413	0.3413	0.1359	0.0228
$n = 2$	$Unif$	0.01648	0.15822	0.32610	0.32558	0.15716	0.01646
	Exp	0.00000	0.11656	0.47684	0.26162	0.09878	0.04620
$n = 16$	$Unif$	0.02340	0.13678	0.34134	0.33680	0.13926	0.02242
	Exp	0.00808	0.14872	0.37438	0.31216	0.12328	0.03338
$n = 36$	$Unif$	0.02322	0.13578	0.34104	0.34108	0.13584	0.02304
	Exp	0.01348	0.14330	0.36422	0.32142	0.12668	0.03090
$n = 100$	$Unif$	0.02194	0.13686	0.34072	0.34068	0.13694	0.02286
	Exp	0.01696	0.14242	0.35276	0.32958	0.13036	0.02792

Example 6.14 Suppose that the shelf life, the number of days a product is on a store's shelf, for 1-gallon cartons of milk is a random variable with a $Unif[1, 7]$ distribution. If a store puts out 100 cartons of 1-gallon of milk for sale, find the probability that the average number of days the cartons remain on the shelf exceeds 4.5 days.

Solution: Let the random variable X represent the number of days a 1 gallon carton of milk is on a store's shelf. Since $X \sim Unif[1, 7]$, using the equations from (4.9), the **pdf** of X can be written as

$$f(x) = \frac{1}{6} \quad \text{if } x \in [1, 7],$$

and the mean and variance of X as

$$E[X] = \frac{a+b}{2} = \frac{1+7}{2} = 4, \quad \text{and} \quad Var[X] = \frac{(b-a)^2}{12} = \frac{(7-1)^2}{12} = 3.$$

Let $X_i, i = 1, \dots, 100$, represent the actual times cartons of milk remain on the store's shelf. Since

$$E[X_i] = 4 \quad \text{and} \quad Var[X_i] = 3,$$

the average time is computed as

$$\bar{X} = \frac{1}{100}(X_1 + \dots + X_{100}).$$

Consequently, the mean and variance of this sample mean are

$$E[\bar{X}] = 4, \quad Var[\bar{X}] = \frac{\sigma^2}{n} = \frac{3}{100} = 0.03.$$

Appealing to the Central Limit Theorem, write

$$\frac{\bar{X} - E[\bar{X}]}{\sqrt{Var[\bar{X}]}} = \frac{\bar{X} - 4}{\sqrt{0.03}} \sim N(0, 1),$$

which is equivalent to writing $\bar{X} \sim N(4, \sqrt{0.03})$. Consequently,

$$\mathbb{P}(\bar{X} > 4.5) = \mathbb{P}\left(Z > \frac{4.5 - 4}{\sqrt{0.03}}\right) = \mathbb{P}(Z > 2.89) = 0.002.$$

The following code computes the answer with S:

```
> round(1 - pnorm(4.5, 4, sqrt(.03)), 3)
[1] 0.002
```

■

Example 6.15 A building contractor provides a detailed estimate of his charges by listing the price of all of his material and labor charges to the nearest dollar. Suppose the rounding charge errors can be treated as independent random variables following $Unif[-10, 10]$ distributions. If a recent estimate from the building contractor listed 100 charges, find the maximum error for the contractor's estimate with probability of 0.95.

Solution: Using the equations from (4.9), if $e_i, i = 1, \dots, 100$ are the estimate errors, then $E[e_i] = \frac{b+a}{2} = 0$ and $Var[e_i] = \frac{(b-a)^2}{12} = \frac{400}{12}$. It follows then that $\mu_{\bar{e}} = 0$ and $\sigma_{\bar{e}}^2 = \frac{400}{1200} = \frac{1}{3}$. Because of the relatively large ($n = 100$) sample size, the Central Limit Theorem tells us that the distribution of \bar{e} is approximately normal with mean 0 and standard deviation $\sqrt{\frac{1}{3}}$.

Since the absolute error of the sum of the 100 charges is the sum of each one of the rounded errors, $e = e_1 + \cdots + e_{100}$, $e = \bar{e} \cdot n$. Written mathematically,

$$\mathbb{P} \left(-1.96 < \frac{\bar{e} - 0}{\sqrt{\frac{1}{3}}} < 1.96 \right) = 0.95.$$

Multiplying by $n = 100$ and $\sqrt{\frac{1}{3}}$ gives a probability expression for e :

$$\mathbb{P} \left(\sqrt{\frac{1}{3}} \cdot 100 \cdot (-1.96) < e < \sqrt{\frac{1}{3}} \cdot 100 \cdot (1.96) \right) = 0.95$$

From the last expression, note that the maximum error for the estimate e_{Max} , is $\sqrt{\frac{1}{3}} \cdot 100 \cdot (1.96) = 113.1607$. In other words, the final job will not deviate more than 113 dollars from the original estimate with 95% confidence. ■

6.5.2 Sampling Distribution for $\bar{X} - \bar{Y}$ when Sampling from Two Independent Normal Populations

The sampling distribution for $\bar{X} - \bar{Y}$ is normal with mean $\mu_X - \mu_Y$ and standard deviation $\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}$, where n_X and n_Y are the respective sample sizes. That is,

$$\bar{X} - \bar{Y} \sim N \left(\mu_X - \mu_Y, \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} \right)$$

provided X and Y are independent random variables where $X \sim N(\mu_X, \sigma_X)$ and $Y \sim N(\mu_Y, \sigma_Y)$. Since X and Y are independent normal random variables, the distributions of their means are known. Specifically,

$$\bar{X} \sim N \left(\mu_X, \frac{\sigma_X}{\sqrt{n_X}} \right) \quad \text{and} \quad \bar{Y} \sim N \left(\mu_Y, \frac{\sigma_Y}{\sqrt{n_Y}} \right).$$

Proof: Using the results from Theorem 5.1 on page 176 and letting $X_1 = \bar{X}$, $X_2 = \bar{Y}$, $a_1 = 1$, and $a_2 = -1$, obtain

$$\bar{X} - \bar{Y} \sim N \left(\mu_X - \mu_Y, \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} \right). \quad (6.5)$$

Example 6.16 ▷ **Simulating $\bar{X} - \bar{Y}$** ◁ Use simulation to verify empirically that if $X \sim N(\mu_X, \sigma_X)$ and $Y \sim N(\mu_Y, \sigma_Y)$, the resulting sampling distribution of $\bar{X} - \bar{Y}$ is as given in (6.5). Specifically, generate and store in a vector named `meansX` the means of 1000 samples of size $n_X = 100$ from a normal distribution with $\mu_X = 100$ and $\sigma_X = 10$. Generate and store in a vector named `meansY` the means of 1000 samples of size $n_Y = 81$ from a normal distribution with $\mu_Y = 50$ and $\sigma_Y = 9$. Produce a probability histogram of the differences between `meansX` and `meansY`, and superimpose the probability histogram with a normal density having mean and standard deviation equal to the theoretical mean and standard deviation for $(\bar{X} - \bar{Y})$ in this problem. Compute the mean and standard deviation for the difference between `meansX` and `meansY`. Finally, compute the empirical probability $\mathbb{P}(\bar{X} - \bar{Y} < 52)$ based on the simulated data as well as the theoretical probability $\mathbb{P}(\bar{X} - \bar{Y} < 52)$.

Solution: In the S code that follows, `m` represents the number of samples, and `nx`, `mux`, `sigx`, `ny`, `muy`, `sigy`, `muxy`, `meansX`, `meansY`, and `XY` represent n_X , μ_X , σ_X , n_Y , μ_Y , σ_Y , $\mu_X - \mu_Y$, \bar{X} , \bar{Y} , and $\bar{X} - \bar{Y}$, respectively. The `set.seed()` command is used so the same values can be generated at a later date. Before running the simulation, note that the theoretical distribution $(\bar{X} - \bar{Y}) \sim N(100 - 50 = 50, \sqrt{10^2/100 + 9^2/81} = \sqrt{2})$. The probability histogram for the empirical distribution of $(\bar{X} - \bar{Y})$ is shown in Figure 6.6 on the next page. Note that the empirical mean and standard deviation for $(\bar{X} - \bar{Y})$ are 50.01 and 1.44, respectively, which are very close to the theoretical values of 50 and $\sqrt{2} \approx 1.41$. The empirical probability $\mathbb{P}(\bar{X} - \bar{Y} < 52)$ is computed by determining the proportion of $(\bar{X} - \bar{Y})$ values that are less than 52. Note that the empirical answer for $\mathbb{P}(\bar{X} - \bar{Y} < 52)$ is 0.918, which is in agreement with the theoretical answer to two decimal places.

```
> set.seed(4)
> m <- 1000
> nx <- 100
> ny <- 81
> mux <- 100
> sigx <- 10
> muy <- 50
> sigy <- 9
> muxy <- mux - muy
> sigxy <- sqrt((sigx^2/nx) + (sigy^2/ny))
> meansX <- array(0, m)      # Array of m zeros
> meansY <- array(0, m)      # Array of m zeros
> for(i in 1:m) {meansX[i] <- mean(rnorm(nx, mux, sigx))}
> for(i in 1:m) {meansY[i] <- mean(rnorm(ny, muy, sigy))}
> XY <- meansX - meansY
> ll <- muxy - 3.4 * sigxy
> ul <- muxy + 3.4 * sigxy
> hist(XY, prob = TRUE, xlab = "xbar-ybar", nclass = "scott", col = 13,
+ xlim = c(ll, ul), ylim = c(0, 0.3), main="", ylab="")
> lines(seq(ll, ul, 0.05), dnorm(seq(ll, ul, 0.05), muxy, sigxy), lwd = 3)
> print(round(c(mean(XY), sqrt(var(XY))), 2))
[1] 50.01  1.44
> sum(XY < 52)/1000
[1] 0.918
> round(pnorm(52, 50, sqrt(2)), 2)
[1] 0.92
```

■

6.5.3 Sampling Distribution for the Sample Proportion

When Y is a binomial random variable, $Y \sim \text{Bin}(n, \pi)$, that represents the number of successes obtained in n trials where the probability of success is π , the sample proportion of successes is typically computed as

$$P = \frac{Y}{n}. \quad (6.6)$$

The mean and variance, respectively, of the sample proportion of successes are

$$E[P] = \mu_P = \pi \quad (6.7)$$

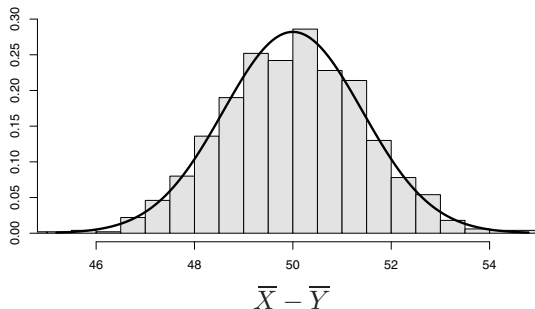


FIGURE 6.6: Probability histogram for simulated distribution of $(\bar{X} - \bar{Y})$ with superimposed normal density with $\mu = 50$ and $\sigma = \sqrt{2}$.

and

$$\text{Var}[P] = \sigma_P^2 = \frac{\pi(1-\pi)}{n}. \quad (6.8)$$

Equations (6.7) and (6.8) are easily derivable using the mean and variance of Y . Since

$$E[Y] = n\pi \quad \text{and} \quad \text{Var}[Y] = n\pi(1-\pi),$$

it follows that

$$E[P] = E\left[\frac{Y}{n}\right] = \frac{1}{n}E[Y] = \pi,$$

and

$$\sigma_P^2 = \text{Var}[P] = \text{Var}\left[\frac{Y}{n}\right] = \frac{1}{n^2}\text{Var}[Y] = \frac{\pi(1-\pi)}{n}.$$

The Central Limit Theorem tells us that the proportion of successes is asymptotically normal for sufficiently large values of n . So that the distribution of P is not overly skewed, both $n\pi$ and $n(1-\pi)$ must be greater than or equal to 5. The larger $n\pi$ and $n(1-\pi)$ are, the closer the distribution of P comes to resembling a normal distribution. The rationale for applying the Central Limit Theorem to the proportion of successes rests on the fact that the sample proportion can also be thought of as a sample mean. Specifically,

$$P = \frac{Y_1 + \cdots + Y_n}{n},$$

where each Y_i value takes on a value of 1 if the element possesses the particular attribute being studied and a 0 if it does not. That is, P is the sample mean for the Bernoulli random variable Y_i . Viewed in this fashion, write

$$Z = \frac{P - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0, 1). \quad (6.9)$$

It is also fairly common to approximate the sampling distribution of Y with a normal distribution using the relationship

$$Z = \frac{Y - n\pi}{\sqrt{n\pi(1-\pi)}} \sim N(0, 1). \quad (6.10)$$

Example 6.17 In plain variety M&M candies, the percentage of green candies is 10%. Suppose a large bag of M&M candies contains 500 candies. What is the probability there will be

- (a) at least 11% green M&Ms?
 (b) no more than 12% green M&Ms?

Solution: First, note that the population proportion of green M&Ms is $\pi = 0.10$. Since neither $n \times \pi = 400 \times 0.10 = 40$ nor $n \times (1 - \pi) = 400 \times 0.90 = 360$ is less than 5, it seems reasonable to appeal to the Central Limit Theorem for the approximate distribution of P . Consequently,

$$P \sim N\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right),$$

which, when using the numbers from the problem, becomes

$$P \sim N\left(0.10, \sqrt{\frac{(0.10)(0.90)}{500}} = 0.01341641\right).$$

If the random variable Y is equal to the number of green M&Ms, then the distribution of Y can be approximated by

$$Y \sim N\left(n\pi, \sqrt{n\pi(1-\pi)}\right),$$

which, when using the numbers from the problem, becomes

$$Y \sim N\left(50, \sqrt{500 \cdot 0.10 \cdot (1 - 0.10)} = 6.708204\right).$$

It is also possible to give the exact distribution of Y , which is $Y \sim \text{Bin}(n = 500, \pi = 0.10)$.

(a) The probabilities that at least 11% of the candies will be green M&Ms using the approximate distribution of P , the approximate distribution of Y , and finally using the exact distribution of Y are as follows:

$$\begin{aligned} \mathbb{P}(P \geq 0.11) &= \mathbb{P}\left(\frac{P - \pi}{\sigma_P} \geq \frac{0.11 - \pi}{\sigma_P}\right) \approx \mathbb{P}\left(Z \geq \frac{0.11 - 0.10}{0.01341641}\right) \\ &= \mathbb{P}(Z \geq 0.745356) = 0.2280283 \\ \mathbb{P}(Y \geq 55) &= \mathbb{P}\left(\frac{Y - n\pi}{\sqrt{n\pi(1-\pi)}} > \frac{55 - n\pi}{\sqrt{n\pi(1-\pi)}}\right) \approx \mathbb{P}\left(Z \geq \frac{55 - 50}{6.708204}\right) \\ &= \mathbb{P}(Z \geq 0.745356) = 0.2280283 \\ \mathbb{P}(Y \geq 55) &= \sum_{i=55}^{500} \binom{500}{i} (0.10)^i (0.90)^{500-i} = 0.2476933 \end{aligned}$$

(b) The probability that no more than 12% of the candies will be green M&Ms is

$$\begin{aligned}\mathbb{P}(P \leq 0.12) &= \mathbb{P}\left(\frac{P - \pi}{\sigma_P} \leq \frac{0.12 - \pi}{\sigma_P}\right) \approx \mathbb{P}\left(Z \leq \frac{0.12 - 0.10}{0.01341641}\right) \\ &= \mathbb{P}(Z \leq 1.490712) = 0.9319814 \\ \mathbb{P}(Y \leq 60) &= \mathbb{P}\left(\frac{Y - n\pi}{\sqrt{n\pi(1-\pi)}} > \frac{60 - n\pi}{\sqrt{n\pi(1-\pi)}}\right) \approx \mathbb{P}\left(Z \leq \frac{60 - 50}{6.708204}\right) \\ &= \mathbb{P}(Z \leq 1.490712) = 0.9319814 \\ \mathbb{P}(Y \leq 60) &= \sum_{i=0}^{60} \binom{500}{i} (0.10)^i (0.90)^{500-i} = 0.9381745.\end{aligned}$$

The following S commands compute the answers for (a) and (b):

```
> 1 - pnorm(0.11,0.10, sqrt(0.1*0.9/500))
[1] 0.2280283
> 1 - pnorm(55,500*0.1, sqrt(500*0.1*0.9))
[1] 0.2280283
> 1 - pbinom(54,500,0.10)
[1] 0.2476933

> pnorm(0.12,0.10, sqrt(0.1*0.9/500))
[1] 0.9319814
> pnorm(60,500*.10, sqrt(500*0.1*0.9))
[1] 0.9319814
> pbinom(60,500,0.1)
[1] 0.9381745
```

The astute observer will notice that the approximations are not equal to the exact answers. This is due to the fact that a continuous distribution has been used to approximate a discrete distribution. The accuracy of the answers can be improved by applying what is called a **continuity correction**. Using the continuity correction, (6.9) and (6.10) become

$$Z = \frac{P \pm \frac{0.5}{n} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \overset{\sim}{\sim} N(0, 1) \quad (6.11)$$

and

$$Z = \frac{Y \pm 0.5 - n\pi}{\sqrt{n\pi(1-\pi)}} \overset{\sim}{\sim} N(0, 1). \quad (6.12)$$

When solving less than or equal type inequalities, add the continuity correction; and when solving greater than or equal type inequalities, subtract the continuity correction. Notice how much closer the approximations are to the exact answers when the appropriate continuity corrections are applied:

$$\begin{aligned}
\mathbb{P}(P \geq 0.11) &= \mathbb{P}\left(\frac{P - \frac{0.5}{500} - \pi}{\sigma_P} \geq \frac{0.11 - \frac{0.5}{500} - \pi}{\sigma_P}\right) \\
&\approx \mathbb{P}\left(Z \geq \frac{0.11 - \frac{0.5}{500} - 0.10}{0.01341641}\right) \\
&= \mathbb{P}(Z \geq 0.6708204) = 0.2511675 \\
\mathbb{P}(Y \geq 55) &= \mathbb{P}\left(\frac{Y - 0.5 - n\pi}{\sqrt{n\pi(1-\pi)}} > \frac{55 - 0.5 - n\pi}{\sqrt{n\pi(1-\pi)}}\right) \\
&\approx \mathbb{P}\left(Z \geq \frac{55 - 0.5 - 50}{6.708204}\right) \\
&= \mathbb{P}(Z \geq 0.6708204) = 0.2511675 \\
\mathbb{P}(Y \geq 55) &= \sum_{i=55}^{500} \binom{500}{i} (0.10)^i (0.90)^{500-i} = 0.2476933
\end{aligned}$$

$$\begin{aligned}
\mathbb{P}(P \leq 0.12) &= \mathbb{P}\left(\frac{P + \frac{0.5}{500} - \pi}{\sigma_P} \leq \frac{0.12 + \frac{0.5}{500} - \pi}{\sigma_P}\right) \\
&\approx \mathbb{P}\left(Z \leq \frac{0.12 + \frac{0.5}{500} - 0.10}{0.01341641}\right) \\
&= \mathbb{P}(Z \leq 1.565248) = 0.9412376 \\
\mathbb{P}(Y \leq 60) &= \mathbb{P}\left(\frac{Y + 0.5 - n\pi}{\sqrt{n\pi(1-\pi)}} > \frac{60 + 0.5 - n\pi}{\sqrt{n\pi(1-\pi)}}\right) \\
&\approx \mathbb{P}\left(Z \leq \frac{60 + 0.5 - 50}{6.708204}\right) \\
&= \mathbb{P}(Z \leq 1.565248) = 0.9412376 \\
\mathbb{P}(Y \leq 60) &= \sum_{i=0}^{60} \binom{500}{i} (0.10)^i (0.90)^{500-i} = 0.9381745
\end{aligned}$$

Example 6.18 The 1999 North Carolina Department of Public Instruction, NC Youth Tobacco Use Survey, reported that 38.3% of all North Carolina high school students used tobacco products. If a random sample of 250 North Carolina high school students is taken, find the probability that the sample proportion that use tobacco products will be between 0.36 and 0.40 inclusive.

Solution: Since neither $n \times \pi = 250 \times 0.383 = 95.75$ nor $n \times (1 - \pi) = 250 \times 0.617 = 154.25$ is less than 5, it seems reasonable to appeal to the Central Limit Theorem for the approximate distribution of P . Consequently,

$$P \rightsquigarrow N\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right),$$

which, when using the numbers from the problem, becomes

$$P \rightsquigarrow N\left(0.383, \sqrt{\frac{(0.383)(0.617)}{250}} = 0.03074482\right).$$

Due to the discrete nature of the problem, appropriate continuity corrections should be used:

$$\mathbb{P}\left(.36 - \frac{0.5}{250} \leq P \leq .40 + \frac{0.5}{250}\right) = \mathbb{P}(.358 \leq P \leq .402) = 0.5236417$$

To calculate $\mathbb{P}(0.358 \leq P \leq 0.402)$ with **S**, use `pnorm()`:

```
> sig <- sqrt((0.383*0.617)/250)
> pnorm(0.402,0.383, sig) - pnorm(0.358,0.383, sig)
[1] 0.5236417
```

The exact answer to the problem can be solved using the binomial distribution as follows:

```
> pbinom(100,250,.383) - pbinom(89,250,.383)
[1] 0.5241166
```

6.5.4 Expected Value and Variance of the Uncorrected Sample Variance and the Sample Variance

Given a random sample X_1, X_2, \dots, X_n taken from a population with mean μ and variance σ^2 , the expected value of the uncorrected variance, S_u^2 , is

$$E[S_u^2] = \frac{1}{n} \sum_{i=1}^n E[(X_i - \bar{X})^2]. \quad (6.13)$$

Expanding the right-hand side of (6.13) gives

$$\begin{aligned} \sum_{i=1}^n (X_i - \bar{X})^2 &= \sum_{i=1}^n [(X_i - \mu) + (\mu - \bar{X})]^2 \\ &= \sum_{i=1}^n \left[(X_i - \mu)^2 + 2(\mu - \bar{X})(X_i - \mu) + (\mu - \bar{X})^2 \right] \\ &= \sum_{i=1}^n (X_i - \mu)^2 + 2(\mu - \bar{X}) \sum_{i=1}^n (X_i - \mu) + n(\mu - \bar{X})^2 \\ &= \sum_{i=1}^n (X_i - \mu)^2 + 2(\mu - \bar{X})(n\bar{X} - n\mu) + n(\mu - \bar{X})^2 \\ &= \sum_{i=1}^n (X_i - \mu)^2 - 2n(\mu - \bar{X})^2 + n(\mu - \bar{X})^2 \\ &= \sum_{i=1}^n (X_i - \mu)^2 - n(\mu - \bar{X})^2. \end{aligned} \quad (6.14)$$

Substituting the expression $\sum_{i=1}^n (X_i - \mu)^2 - n(\mu - \bar{X})^2$ for $\sum_{i=1}^n (X_i - \bar{X})^2$ in (6.13) gives

$$\begin{aligned} E[S_u^2] &= \frac{1}{n} E \left[\sum_{i=1}^n (X_i - \mu)^2 - n(\mu - \bar{X})^2 \right] \\ E[S_u^2] &= \frac{1}{n} \left(n\sigma^2 - n \frac{\sigma^2}{n} \right) \\ E[S_u^2] &= \sigma^2 - \frac{\sigma^2}{n} \\ &= \sigma^2 \left(\frac{n-1}{n} \right). \end{aligned} \quad (6.15)$$

As (6.15) shows, the expected value of S_u^2 , $\sigma^2\left(\frac{n-1}{n}\right)$, is less than σ^2 . However, as n increases, this difference diminishes. The variance for the uncorrected variance S_u^2 , is given by

$$\text{Var}[S_u^2] = \frac{\mu_4 - \mu_2^2}{n} - \frac{2(\mu_4 - 2\mu_2^2)}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3}, \quad (6.16)$$

where $\mu_k = E[(X - \mu)^k]$ is the k^{th} central moment. Using the definition for the sample variance from (6.4), the expected value of S^2 is readily verified to be σ^2 .

The probability distributions for S_u^2 and S^2 are typically skewed to the right. The skewness diminishes as n increases. Of course, the Central Limit Theorem indicates that the distributions of both are asymptotically normal. However, the convergence to a normal distribution is very slow and requires a very large n . The distributions of S_u^2 and S^2 are extremely important in statistical inference. Two special cases, examined next, are the sampling distributions of S_u^2 and S^2 when sampling from normal populations.

6.6 Sampling Distributions Associated with the Normal Distribution

6.6.1 Chi-Square Distribution (χ^2)

The chi-square distribution is a special case of the gamma distribution covered in Section 4.3.3 on page 139. In a paper published in 1900, Karl Pearson popularized the use of the chi-square distribution to measure goodness-of-fit. The **pdf**, $E(X)$, $\text{Var}(X)$, and the **mgf** for a chi-square random variable are given in (6.17), where $\Gamma\left(\frac{n}{2}\right)$ is defined in (4.15).

<p>Chi-Square Distribution $X \sim \chi_n^2$</p> $f(x) = \begin{cases} \frac{1}{\Gamma\left(\frac{n}{2}\right) 2^{\frac{n}{2}}} \cdot x^{\frac{n}{2}-1} e^{-\frac{x}{2}} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \quad (6.17)$ <p> $E[X] = n$ $\text{Var}[X] = 2n$ $M_X(t) = (1 - 2t)^{-\frac{n}{2}}$ for $t < \frac{1}{2}$ </p>
--

The chi-square distribution is strictly dependent on the parameter n , called the **degrees of freedom**. In general, the chi-square distribution is unimodal and skewed to the right. Three different chi-square distributions are represented in Figure 6.7 on the next page. The notation used with the chi-square distribution to indicate α of the distribution is in the left tail when the distribution has n degrees of freedom is $\chi_{\alpha;n}^2$. For example, $\chi_{0.95;10}^2$ denotes the value such that 95% of the area is to the left of said value in a χ_{10}^2 distribution.

To find the value corresponding to $\chi_{0.95;10}^2$, use the **S** command `qchisq(p, df)`, where p is the area to the left (probability) and `df` is the degrees of freedom. The command gives

```
> qchisq(.95, 10)
[1] 18.30704
```

which says that $\mathbb{P}(\chi_{10}^2 < 18.31) = 0.95$.

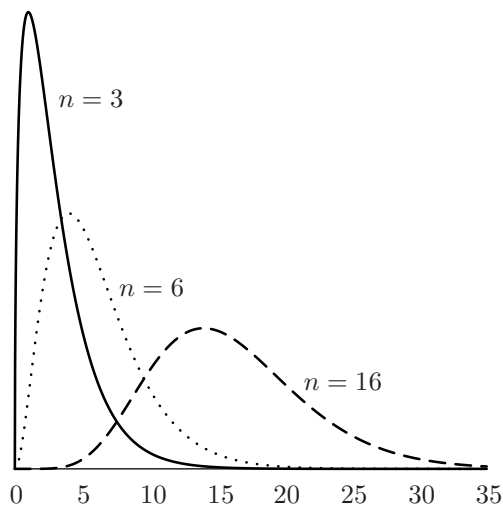


FIGURE 6.7: Illustrations of the pdfs of χ_3^2 , χ_6^2 , and χ_{16}^2 random variables

Asymptotic properties. For large values of n ($n > 100$), the distribution of $\sqrt{2\chi_n^2}$ has an approximate normal distribution with a mean of $\sqrt{2n-1}$ and a standard deviation of 1. In other words, because $\sqrt{2\chi_n^2} \rightsquigarrow N(\sqrt{2n-1}, 1)$, $Y = \sqrt{2\chi_n^2} - \sqrt{2n-1} \rightsquigarrow N(0, 1)$. For very large values of n , the approximation

$$Y = \frac{\chi_n^2 - n}{\sqrt{2n}} \rightsquigarrow N(0, 1)$$

may also be used.

Example 6.19 Compute the indicated quantities:

- $\mathbb{P}(\chi_{150}^2 \geq 126)$
- $\mathbb{P}(40 \leq \chi_{65}^2 \leq 50)$
- $\mathbb{P}(\chi_{220}^2 \geq 260)$
- The value a such that $\mathbb{P}(\chi_{100}^2 \leq a) = 0.6$

Solution: The answers are computed first by hand using the approximation $\sqrt{2\chi_n^2} \rightsquigarrow N(\sqrt{2n-1}, 1)$. Then, the exact probabilities are calculated with S.

$$(a) \mathbb{P}(\chi_{150}^2 \geq 126) = \mathbb{P}(\sqrt{2\chi_{150}^2} - \sqrt{299} \geq \sqrt{2(126)} - \sqrt{299}) \approx \mathbb{P}(Z \geq -1.42) = 0.922.$$

```
> 1 - pchisq(126, 150)
[1] 0.923393
```

(b)

$$\begin{aligned}
\mathbb{P}(40 \leq \chi_{65}^2 \leq 50) &= \mathbb{P}(\sqrt{2(40)} \leq \sqrt{2\chi_{65}^2} \leq \sqrt{2(50)}) \\
&= \mathbb{P}(\sqrt{80} - \sqrt{129} \leq \sqrt{2\chi_{65}^2} - \sqrt{129} \leq \sqrt{100} - \sqrt{129}) \\
&\approx \mathbb{P}(-2.41 \leq Z \leq -1.36) = 0.079.
\end{aligned}$$

```
> pchisq(50,65) - pchisq(40,65)
[1] 0.07861696
```

(c)

$$\begin{aligned}
\mathbb{P}(\chi_{220}^2 \geq 260) &= \mathbb{P}(\sqrt{2\chi_{220}^2} \geq \sqrt{2 \cdot 260}) \\
&= \mathbb{P}(\sqrt{2\chi_{220}^2} - \sqrt{2(220) - 1} \geq \sqrt{2 \cdot 260} - \sqrt{2(220) - 1}) \\
&\approx \mathbb{P}(Z \geq 1.85) = 0.032.
\end{aligned}$$

```
> 1 - pchisq(260,220)
[1] 0.03335803
```

(d)

$$\begin{aligned}
&\mathbb{P}(\chi_{100}^2 \leq a) = 0.6 \\
\mathbb{P}\left(\sqrt{2\chi_{100}^2} - \sqrt{2(100) - 1} \leq \sqrt{2a} - \sqrt{2(100) - 1}\right) &= 0.6 \\
\mathbb{P}\left(Z \leq \sqrt{2a} - \sqrt{2(100) - 1}\right) &= 0.6 \\
0.2533 = \sqrt{2a} - \sqrt{199} & \\
\Rightarrow a = 103.106. &
\end{aligned}$$

```
> qchisq(.6,100)
[1] 102.9459
```

Note that the approximations are close to the answers from S, but they are not exactly equal. ■

6.6.1.1 The Relationship between the χ^2 Distribution and the Normal Distribution

In addition to describing the χ^2 distribution as a special case of the gamma distribution, the χ^2 distribution can be defined as the sum of independent, squared, standard normal random variables. If n is the number of summed independent, squared, standard normal random variables, then the resulting distribution is a χ^2 distribution with n degrees of freedom, written χ_n^2 . That is,

$$\chi_n^2 = \sum_{i=1}^n Z_i^2, \quad Z_i \sim N(0, 1). \quad (6.18)$$

To complete the proof of Theorem 6.1, recall that the derivative inside the integral when certain characteristics are satisfied is

$$\frac{d}{d\theta} \int_{a(\theta)}^{b(\theta)} f(x, \theta) dx = f(b(\theta), \theta) \frac{d}{d\theta} b(\theta) - f(a(\theta), \theta) \frac{d}{d\theta} a(\theta) + \int_{a(\theta)}^{b(\theta)} \frac{\partial f(x, \theta)}{\partial \theta} dx.$$

For the proof, the integral needed is

$$\begin{aligned} \frac{d}{d\theta} \int_{a(\theta)}^{b(\theta)} f(x; \theta) dx &= \frac{d}{dy} \int_0^{\sqrt{y}} e^{-x^2/2} dx \\ &= f(\sqrt{y}) \frac{d}{dy} (\sqrt{y}) - f(0) \frac{d}{dy} (0) + \int_0^{\sqrt{y}} \frac{\partial e^{-x^2/2}}{\partial y} dx \\ &= e^{-y/2} \frac{1}{2\sqrt{y}}. \end{aligned}$$

Theorem 6.1 If $Z \sim N(0, 1)$, then the random variable $Y = Z^2 \sim \chi_1^2$.

Proof: In this proof, it is shown that the distribution of Y is a χ_1^2 :

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Z^2 \leq y) = \mathbb{P}(-\sqrt{y} \leq Z \leq \sqrt{y}) \\ &= 2\mathbb{P}(0 \leq Z \leq \sqrt{y}) = \frac{2}{\sqrt{2\pi}} \int_0^{\sqrt{y}} e^{-x^2/2} dx \end{aligned}$$

Taking the derivative of $F_Y(y)$ yields

$$f(y) = \frac{dF_Y(y)}{dy} = \frac{2}{\sqrt{2\pi}} \frac{1}{2\sqrt{y}} e^{-y/2} = \frac{1}{\sqrt{2}\Gamma(1/2)} y^{(1/2)-1} e^{-y/2}, \quad 0 \leq y < \infty,$$

which is the **pdf** for a χ_1^2 .

Corollary 6.1 If $X \sim N(\mu, \sigma)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$, and $Z^2 \sim \chi_1^2$.

Theorem 6.2 If X_1, \dots, X_r are independent random variables with chi-square distributions $\chi_{n_1}^2, \dots, \chi_{n_r}^2$, respectively, then

$$Y = \sum_{i=1}^r X_i \sim \chi_s^2, \quad \text{where } s = \sum_{i=1}^r n_i.$$

Proof:

$$\begin{aligned} M_Y(t) &= E[e^{tY}] = \prod_{i=1}^r E[e^{tX_i}] = \prod_{i=1}^r M_{X_i}(t) \\ &= \prod_{i=1}^r (1 - 2t)^{-\frac{n_i}{2}} = (1 - 2t)^{-\frac{1}{2} \sum_{i=1}^r n_i} \end{aligned}$$

which is the **mgf** for a χ_s^2 distribution.

One of the properties of χ^2 distributions is that of reproducibility. In other words, the sum of independent χ^2 random variables is also a χ^2 distribution with degrees of freedom equal to the sum of the degrees of freedom of each of the independent χ^2 random variables.

Corollaries 6.2 and 6.3 are direct consequences of Theorem 6.2 on the preceding page.

Corollary 6.2 If X_1, \dots, X_n are independent random variables following a $N(0, 1)$ distribution, then

$$Y = \sum_{i=1}^n X_i^2 \sim \chi_n^2.$$

Corollary 6.3 If X_1, \dots, X_n are independent random variables with $N(\mu_i, \sigma_i)$ distributions, respectively, then

$$Y = \sum_{i=1}^n \frac{(X_i - \mu_i)^2}{\sigma_i^2} \sim \chi_n^2.$$

Example 6.20 Given 10 independent and identically distributed (i.i.d.) random variables Y_i , where $Y_i \sim N(0, \sigma = 5)$ for $i = 1, \dots, 10$, compute

(a) $\mathbb{P}\left(\sum_{i=1}^{10} Y_i^2 \leq 600\right)$

(b) $\mathbb{P}\left(\frac{1}{10} \sum_{i=1}^{10} Y_i^2 \geq 12.175\right)$

(c) The number a such that $\mathbb{P}\left(\sqrt{\frac{1}{10} \sum_{i=1}^{10} Y_i^2} \geq a\right) = 0.5$

Solution: The answers are computed using `S`. Be sure to note that $Z = \frac{Y_i - 0}{5} = \frac{Y_i}{5}$.

$$\begin{aligned} \mathbb{P}\left(\sum_{i=1}^{10} Y_i^2 \leq 600\right) &= \mathbb{P}\left(\sum_{i=1}^{10} \left(\frac{Y_i}{5}\right)^2 \leq \frac{600}{25}\right) \\ &= \mathbb{P}(\chi_{10}^2 \leq 24) > 0.99. \end{aligned}$$

Using the `S` command `pchisq(24, 10)` gives $\mathbb{P}(\chi_{10}^2 \leq 24) = 0.9923996$:

```
> pchisq(24, 10)
[1] 0.9923996
```

(b)

$$\begin{aligned} \mathbb{P}\left(\frac{1}{10} \sum_{i=1}^{10} Y_i^2 \geq 12.175\right) &= \mathbb{P}\left(\sum_{i=1}^{10} \left(\frac{Y_i}{5}\right)^2 \geq \frac{12.175(10)}{25}\right) \\ &= \mathbb{P}(\chi_{10}^2 \geq 4.87) = 0.90. \end{aligned}$$

```
> 1 - pchisq(4.87, 10)
[1] 0.8996911
```

(c)

$$\begin{aligned}\mathbb{P}\left(\frac{1}{10}\sum_{i=1}^{10}Y_i^2 \geq a^2\right) &= \mathbb{P}\left(\sum_{i=1}^{10}\left(\frac{Y_i}{5}\right)^2 \geq \frac{10a^2}{25}\right) \\ &= \mathbb{P}\left(\chi_{10}^2 \geq \frac{10a^2}{25}\right) = 0.5\end{aligned}$$

Using the S command `qchisq()`, the value $\chi_{10,0.50}^2 = 9.34$ is calculated:

```
> qchisq(0.50,10)
[1] 9.341818
```

Consequently, $\frac{10a^2}{25} = 9.34$, which yields $a = 4.83$. ■

6.6.1.2 Sampling Distribution for S_u^2 and S^2 when Sampling from Normal Populations

In this section, the resulting sampling distributions for S_u^2 and S^2 given in Table 6.4 on page 205 when sampling from a normal distribution are considered. Note that $\sum_{i=1}^n (X_i - \bar{X})^2 = nS_u^2 = (n-1)S^2$ and that dividing this by σ^2 yields

$$\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2} = \frac{nS_u^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} \quad (6.19)$$

The first term in (6.19) appears to be some type of standardized normal random variable. However, it is not, since the sample mean of a random variable is itself a random variable and not a constant. So, what is the distribution then of nS_u^2/σ^2 ? Theorem 6.3 tells us that the distribution of nS_u^2/σ^2 is χ_{n-1}^2 .

Theorem 6.3 Let X_1, \dots, X_n be a random sample from a $N(\mu, \sigma)$ distribution. Then,

- (1) \bar{X} and S^2 are independent random variables. Likewise, \bar{X} and S_u^2 are independent random variables.
- (2) The random variable

$$\frac{nS_u^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi_{n-1}^2$$

Proof: A detailed proof of part (1) in Theorem 6.3 is beyond the scope of the text, and the statement will simply be assumed to be true. The independence between \bar{X} and S^2 is a result of normal distributions. Almost without exception, the estimators \bar{X} and S^2 are dependent in all other distributions.

To prove part (2) of Theorem 6.3, use Corollary 6.3 to say that $\sum_{i=1}^n \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi_n^2$. Then, rearrange the terms to find an expression for $\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2}$ for which the distribution is recognizable. Start by rearranging the numerator of the χ_n^2 distribution:

$$\begin{aligned}\sum_{i=1}^n (X_i - \mu)^2 &= \sum_{i=1}^n [(X_i - \bar{X}) + (\bar{X} - \mu)]^2 \\ &= \sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^n (\bar{X} - \mu)^2 + 2 \sum_{i=1}^n (X_i - \bar{X})(\bar{X} - \mu)\end{aligned}$$

Since

$$\sum_{i=1}^n (X_i - \bar{X}) (\bar{X} - \mu) = (\bar{X} - \mu) \sum_{i=1}^n (X_i - \bar{X}) = 0,$$

it follows that

$$\sum_{i=1}^n (X_i - \mu)^2 = \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2. \quad (6.20)$$

Dividing (6.20) by σ^2 gives

$$\frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} + \frac{n(\bar{X} - \mu)^2}{\sigma^2},$$

which is the same as

$$\frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} + \frac{n(\bar{X} - \mu)^2}{\sigma^2}. \quad (6.21)$$

Since $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$, it follows that $\frac{n(\bar{X} - \mu)^2}{\sigma^2} \sim \chi_1^2$ by Corollary 6.1 on page 229. To simplify notation, let Y , Y_1 , and Y_2 represent $\frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^2}$, $\frac{(n-1)S^2}{\sigma^2}$, and $\frac{n(\bar{X} - \mu)^2}{\sigma^2}$ in (6.21), respectively. By part (1) of Theorem 6.3 on the preceding page, Y_1 and Y_2 are independent. Therefore,

$$\begin{aligned} E[e^{tY}] &= E[e^{t(Y_1 + Y_2)}] = E[e^{tY_1}] \cdot E[e^{tY_2}] \\ (1 - 2t)^{-\frac{n}{2}} &= E[e^{tY_1}] \cdot (1 - 2t)^{-\frac{1}{2}} \\ (1 - 2t)^{-\frac{(n-1)}{2}} &= E[e^{tY_1}] = M_{Y_1}(t) \Rightarrow Y_1 \sim \chi_{n-1}^2. \end{aligned}$$

Note that $Y_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \chi_{n-1}^2$ is based on the n quantities $X_1 - \bar{X}, X_2 - \bar{X}, \dots, X_n - \bar{X}$, which sum to zero. Consequently, specifying the values of any $n - 1$ of the quantities determines the remaining value. That is, only $n - 1$ of the quantities are free to vary. In contrast, $Y = \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^2} \sim \chi_n^2$ has n degrees of freedom since there are no restrictions on the quantities $X_1 - \mu, X_2 - \mu, \dots, X_n - \mu$. In general, when statistics are used to estimate parameters, one degree of freedom is lost for each estimated parameter.

Example 6.21 Show that $E(S_u^2)$, $E(S^2)$, $Var(S_u^2)$, and $Var(S^2)$ are equal to $\frac{(n-1)\sigma^2}{n}$, σ^2 , $\frac{2(n-1)\sigma^4}{n^2}$, and $\frac{2\sigma^4}{n-1}$, respectively, when sampling from a normal distribution.

Solution: It is known that $\frac{nS_u^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$ according to Theorem 6.3 on the previous page. Therefore,

(a)

$$\begin{aligned} E\left[\frac{nS_u^2}{\sigma^2}\right] &= E[\chi_{n-1}^2] = n - 1, \text{ so} \\ \frac{n}{\sigma^2} E[S_u^2] &= n - 1 \Rightarrow E[S_u^2] = \frac{(n-1)\sigma^2}{n} \end{aligned}$$

(b)

$$\begin{aligned} E\left[\frac{(n-1)S^2}{\sigma^2}\right] &= E[\chi_{n-1}^2] = n - 1 \\ \frac{(n-1)}{\sigma^2} E[S^2] &= n - 1 \Rightarrow E[S^2] = \sigma^2 \end{aligned}$$

(c)

$$\begin{aligned} \text{Var} \left[\frac{nS_u^2}{\sigma^2} \right] &= \text{Var} [\chi_{n-1}^2] = 2(n-1) \\ \frac{n^2}{\sigma^4} \text{Var} [S_u^2] &= 2(n-1) \Rightarrow \text{Var} [S_u^2] = \frac{2(n-1)\sigma^4}{n^2} \end{aligned}$$

(d)

$$\begin{aligned} \text{Var} \left[\frac{(n-1)S^2}{\sigma^2} \right] &= \text{Var} [\chi_{n-1}^2] = 2(n-1) \\ \frac{(n-1)^2}{\sigma^4} \text{Var} [S^2] &= 2(n-1) \Rightarrow \text{Var} [S^2] = \frac{2\sigma^4}{(n-1)} \end{aligned}$$

Example 6.22 A random sample of size 11 is taken from a $N(\mu, \sigma)$ distribution where both the mean and the standard deviation are unknown and the sample variance S^2 is computed. Compute the $\mathbb{P}(0.487 < \frac{S^2}{\sigma^2} < 1.599)$.

Solution: According to Theorem 6.3 on page 231, $\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$, which implies $\frac{10S^2}{\sigma^2} \sim \chi_{10}^2$:

$$\begin{aligned} \mathbb{P} \left(0.487 < \frac{S^2}{\sigma^2} < 1.599 \right) &= \mathbb{P} \left(0.487(10) < \frac{10S^2}{\sigma^2} < 1.599(10) \right) \\ &= \mathbb{P}(4.87 < \chi_{10}^2 < 15.99) \\ &= \mathbb{P}(\chi_{10}^2 < 15.99) - \mathbb{P}(\chi_{10}^2 < 4.87) \\ &= 0.90 - 0.10 = 0.80 \end{aligned}$$

To find $\mathbb{P}(\chi_{10}^2 < 15.99)$ and $\mathbb{P}(\chi_{10}^2 < 4.87)$, one can use the **S** command `pchisq()`:

```
> pchisq(15.99,10) - pchisq(4.87,10)
[1] 0.7997721
```

Example 6.23 A custom door manufacturer knows that the measurement error in the height of his final products (the door height minus the order height) follows a normal distribution with a variance of $\sigma^2 = 225 \text{ mm}^2$. A local contractor building custom bungalows orders 31 doors. What is the $\mathbb{P}(S > 18.12 \text{ mm})$ for the 31 doors, and what is the expected value of S^2 ?

Solution:

$$\mathbb{P}(S > 18.12) = \mathbb{P} \left(\frac{n-1}{\sigma^2} S^2 > \frac{30}{225} 18.12^2 \right) = \mathbb{P}(\chi_{30}^2 > 43.78) \approx 0.05$$

The following computes $\mathbb{P}(\chi_{30}^2 > 43.78)$ with **S**:

```
> 1 - pchisq(43.78,30)
[1] 0.04992715
```

Since the expected value of S^2 is the population variance, $E[S^2] = 225$.

Example 6.24 ▷ *Probability Distribution of $(n-1)S^2/\sigma^2$* ◁ Use simulation to generate $m = 1000$ samples of size $n = 15$ from both a $N(0, 1)$ distribution and an $Exp(1)$ distribution. Compute the statistic $(n-1)S^2/\sigma^2$ for both the normally and exponentially generated values, labeling the first NC14 and the second EC14. Produce probability histograms for NC14 and EC14 and superimpose the theoretical distribution for a χ_{14}^2 distribution on both. Repeat the entire process with samples of size $n = 100$. That is, use simulation to generate $m = 1000$ samples of size $n = 100$ from both a $N(0, 1)$ distribution and an $Exp(1)$ distribution. Compute the statistic $(n-1)S^2/\sigma^2$ for both the normally and exponentially generated values, labeling the first NC99 and the later EC99. Produce probability histograms for NC99 and EC99, and superimpose the theoretical distribution for a χ_{99}^2 distribution on both. What can be concluded about the probability distribution of $(n-1)S^2/\sigma^2$ when sampling from a normal distribution and when sampling from an exponential distribution based on the probability histograms?

Solution: The S code that follows generates the required values. To obtain reproducible values, use `set.seed()`. In this solution, `set.seed(302)` is used.

```
> set.seed(302)
> par(mfrow=c(2,2))
> m <- 1000; n <- 15
> varNC14 <- array(0, m)          # Array with m zeros
> for (i in 1:m) {varNC14[i] <- var(rnorm(n))}
> NC14 <- (n-1)*varNC14/1
> hist(NC14, prob=TRUE, ylim=c(0,0.09), xlab="NC14", col=2, xlim=c(0,60),
+ nclass="scott", main="", ylab="")
> lines(seq(0,60,.1), dchisq(seq(0,60,.1), n-1), lwd=3)
> varEC14 <- array(0, m)
> for (i in 1:m) {varEC14[i] <- var(rexp(n))}
> EC14 <- (n-1)*varEC14/1
> hist(EC14, prob=TRUE, ylim=c(0,0.09), xlab="EC14", col=4, xlim=c(0,60),
+ nclass="scott", main="", ylab="")
> lines(seq(0,60,.1), dchisq(seq(0,60,.1), n-1), lwd=3)
> n <- 100
> varNC99 <- array(0, m)
> for (i in 1:m) {varNC99[i] <- var(rnorm(n))}
> NC99 <- (n-1)*varNC99/1
> hist(NC99, prob=TRUE, ylim=c(0,0.03), xlab="NC99", col=2, xlim=c(0,210),
+ nclass="scott", main="", ylab="")
> lines(seq(0,210,.1), dchisq(seq(0,210,.1), n-1), lwd=3)
> varEC99 <- array(0, m)
> for (i in 1:m) {varEC99[i] <- var(rexp(n))}
> EC99 <- (n-1)*varEC99/1
> hist(EC99, prob=TRUE, ylim=c(0,0.03), xlab="EC99", col=4, xlim=c(0,210),
+ nclass="scott", main="", ylab="")
> lines(seq(0,210,.1), dchisq(seq(0,210,.1), n-1), lwd=3)
> NC14 <- c(mean(varNC14), var(varNC14), mean(NC14), var(NC14))
> EC14 <- c(mean(varEC14), var(varEC14), mean(EC14), var(EC14))
> NC99 <- c(mean(varNC99), var(varNC99), mean(NC99), var(NC99))
> EC99 <- c(mean(varEC99), var(varEC99), mean(EC99), var(EC99))
> MAT <- round(rbind(NC14, EC14, NC99, EC99), 4)
> colNAM <- c("E(S^2)", "Var(S^2)", "E(X^2)", "Var(X^2)")
> rowNAM <- c("NC14", "EC14", "NC99", "EC99")
```

```
> dimnames(MAT) <- list(rowNAM , colNAM)
> print(MAT) # Numerical values for Table 6.14
```

Table 6.14: Output for Example 6.24

	$E[S^2]$	$Var[S^2]$	$E\left[\frac{(n-1)S^2}{\sigma^2}\right]$	$Var\left[\frac{(n-1)S^2}{\sigma^2}\right]$
NC14	1.0003	0.1458	14.0039	28.5763
EC14	1.0119	0.5470	14.1666	107.2084
NC99	0.9995	0.0193	98.9491	189.1410
EC99	1.0092	0.0879	99.9125	861.1833

Examine Table 6.14, and note that the means for the simulated S^2 values ($E(S^2)$) for NC14, EC14, NC99, and EC99 are all close to the theoretical variance ($\sigma^2 = 1$). However, only when sampling from a normal distribution does the variance of S^2 equal $2\sigma^4/(n-1)$. That is, the simulated $Var(S^2)$ values for NC14 and NC99 are 0.1458 and 0.0193, which are close to the theoretical values of $2/14 = 0.1428571$ and $2/99 = 0.02020202$. The means and variances for the simulated $(n-1)S^2/\sigma^2$ values are approximately $(n-1)$ and $2(n-1)$, respectively, for NC14 and NC99. However, the variances of $(n-1)S^2/\sigma^2$ when sampling from an exponential are not close to the values returned with NC14 and NC99, nor is the simulated sampling distribution for $(n-1)S^2/\sigma^2$ approximated very well with a χ_{n-1}^2 distribution when sampling from an exponential distribution, as evidenced by the graphs on the right-hand side of Figure 6.8 on the following page. In other words, the sampling distribution for $(n-1)S^2/\sigma^2$ can only be guaranteed to follow a χ_{n-1}^2 distribution when sampling is from a normal distribution. ■

6.6.2 *t*-Distribution

Given a random sample X_1, \dots, X_n that is drawn from a $N(\mu, \sigma)$ distribution, $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$, which implies

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1). \quad (6.22)$$

The quantity (6.22) is used primarily for inference regarding μ . However, this inference assumes σ is known. The assumption of a known σ is generally not reasonable. That is, if μ is unknown, it almost certainly follows that σ will be unknown as well. Fortunately, inference regarding μ can still be performed if σ is replaced by S in (6.22). Specifically, the quantity

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \quad (6.23)$$

follows a well-known distribution, described next.

DEFINITION 6.1: Given two independent random variables Z and U , where $Z \sim N(0, 1)$ and $U \sim \chi_{\nu}^2$, we define the *t*-distribution with ν degrees of freedom as the ratio of Z divided by the square root of U divided by its degrees of freedom. That is,

$$T = \frac{Z}{\sqrt{\frac{U}{\nu}}} \sim t_{\nu}. \quad (6.24)$$

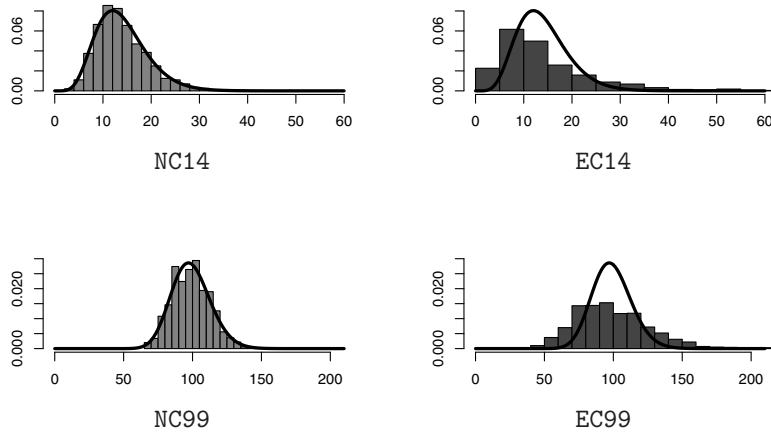


FIGURE 6.8: Probability histograms for simulated distributions of $\frac{(n-1)S^2}{\sigma^2}$ when sampling from normal and exponential distributions. NC14 designates the simulated sampling distributions of $\frac{(n-1)S^2}{\sigma^2}$ when taking samples of sizes $n = 15$ from a normal distribution. In a similar fashion, NC99 denotes the simulated sampling distribution when taking samples of size $n = 100$ from a normal distribution. EC14 and EC99 are analogous to NC14 and NC99 with the exception that the sampling is done from an exponential distribution. The superimposed density on all curves is a χ_{n-1}^2 .

Using definition 6.1, one can readily see why (6.23) follows a t -distribution with $n - 1$ degrees of freedom since

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}} \sqrt{\frac{(n-1)S^2}{(n-1)\sigma^2}}} = \frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{\chi_{n-1}^2}{n-1}}} = \frac{Z}{\sqrt{\frac{U_{n-1}}{n-1}}} \sim t_{n-1}.$$

The t -distribution, also called Student's t -distribution, was first described in a paper published by William Sealy Gosset under the pseudonym "Student." Gosset was employed by Guinness Breweries when his research relating to the t -distribution was published. Since Guinness Breweries had a policy preventing research publications by its staff, Gosset published his findings under the pseudonym "Student." Consequently, the t -distribution is often called Student's t -distribution in his honor. The pdf, expectation, and variance of a t -distribution with ν degrees of freedom are given in (6.25).

t-Distribution		
$X \sim t_\nu$		
$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$	for $-\infty < x < \infty$	(6.25)
$E[X] = 0$		
$Var[X] = \frac{\nu}{\nu - 2}$ for $\nu > 2$		

The shape of the t -distribution is similar to that of the normal distribution; but for small sample sizes, it has heavier tails than the $N(0, 1)$. Figure 6.9 illustrates the densities for t -distributions with 1, 3, and ∞ degrees of freedom, respectively. Note that $t_{\alpha; \infty} = z_{\alpha}$. To find the quantity $t_{\alpha; \nu}$, the S command `pt(α, ν)` can be used. In particular, suppose $t_{0.80; 1}$, depicted in Figure 6.9, is desired. Using the S command `pt(0.80, 1)` gives 1.376382 for the answer.

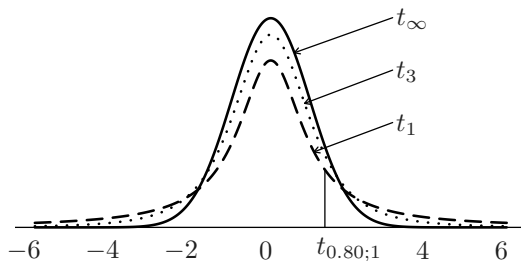


FIGURE 6.9: Illustrations of the pdfs of t_1 (dashed line), t_3 (dotted line), and t_{∞} (solid line) random variables.

Example 6.25 The tensile strength for a type of wire is normally distributed with an unknown mean μ and an unknown variance σ^2 . Five pieces of wire are randomly selected from a large roll, and the strength of each segment of wire is measured. Find the probability that \bar{Y} will be within $\frac{2S}{\sqrt{n}}$ of the true population mean, μ .

Solution: The solution is

$$\begin{aligned} \mathbb{P}\left(\mu - \frac{2S}{\sqrt{n}} \leq \bar{Y} \leq \mu + \frac{2S}{\sqrt{n}}\right) &= \mathbb{P}\left(-\frac{2S}{\sqrt{n}} \leq \bar{Y} - \mu \leq \frac{2S}{\sqrt{n}}\right) \\ &= \mathbb{P}\left(-2 \leq \frac{\bar{Y} - \mu}{S/\sqrt{n}} \leq 2\right) \\ &= \mathbb{P}(-2 \leq t_4 \leq 2) = 0.8838835. \end{aligned}$$

Note that if σ were known, $\mathbb{P}(-2 \leq Z \leq 2) = 0.9544$. ■

The Sampling Distribution for $\bar{X} - \bar{Y}$ when σ_X and σ_Y Are Unknown but Assumed Equal

Theorem 6.4 Given two random samples X_1, \dots, X_{n_X} and Y_1, \dots, Y_{n_Y} that are taken from independent normal populations where $X \sim N(\mu_X, \sigma_X)$, $Y \sim N(\mu_Y, \sigma_Y)$, and $\sigma_X = \sigma_Y$, the random variable

$$\frac{[(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)]}{\sqrt{\frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2} \left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}} \sim t_{n_X + n_Y - 2}. \quad (6.26)$$

Proof: Since $\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}\right)$, according to Theorem 5.1 on page 176,

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \sim N(0, 1).$$

By Theorem 6.3 on page 231, $\frac{(n_X-1)S_X^2}{\sigma_X^2} \sim \chi_{n_X-1}^2$ and $\frac{(n_Y-1)S_Y^2}{\sigma_Y^2} \sim \chi_{n_Y-1}^2$. Since X and Y are independent, it follows that

$$W = \frac{(n_X - 1)S_X^2}{\sigma_X^2} + \frac{(n_Y - 1)S_Y^2}{\sigma_Y^2} \sim \chi_{n_X+n_Y-2}^2$$

from Theorem 6.2 on page 229. Using the definition of the t -distribution, given in definition 6.1 on page 235, $\frac{Z}{\sqrt{\frac{W}{\nu}}} \sim t_\nu$. In this particular case, $\nu = n_X + n_Y - 2$ and, since $\sigma_X = \sigma_Y = \sigma$ is assumed,

$$\begin{aligned} \frac{Z}{\sqrt{\frac{W}{\nu}}} &= \frac{\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}}{\sqrt{\frac{\frac{(n_X - 1)S_X^2}{\sigma_X^2} + \frac{(n_Y - 1)S_Y^2}{\sigma_Y^2}}{n_X + n_Y - 2}}} \\ &= \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sigma \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \cdot \frac{1}{\frac{1}{\sigma} \sqrt{\frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2}}} \\ &= \frac{[(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)]}{\sqrt{\frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2}} \left(\frac{1}{n_X} + \frac{1}{n_Y}\right)} \sim t_{n_X+n_Y-2}. \end{aligned}$$

6.6.3 The F Distribution

In Section 6.6.2, it was seen how the t -distribution can be used to make statements about an unknown mean μ when σ is also unknown. Another common problem statisticians face is that of comparing unknown variances, for example, in manufacturing processes, in mixtures, or in quality from different suppliers of goods. The distribution that allows us to make these comparisons is the F distribution.

DEFINITION 6.2: If U and V are independent random variables, each with a χ^2 distribution with ν_1 and ν_2 degrees of freedom, respectively, then

$$\frac{\frac{U}{\nu_1}}{\frac{V}{\nu_2}} \sim F_{\nu_1, \nu_2}.$$

The **pdf**, expected value, and variance of an F distribution are given in (6.27).

F Distribution

$X \sim F_{\nu_1, \nu_2}$

$$f(x) = \frac{\Gamma\left(\frac{\nu_1 + \nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} x^{\frac{\nu_1}{2} - 1} \left(1 + \frac{\nu_1}{\nu_2}x\right)^{-\frac{1}{2}(\nu_1 + \nu_2)} \quad (6.27)$$

$$E[X] = \frac{\nu_2}{\nu_2 - 2}$$

$$\text{Var}[X] = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)} \text{ provided } \nu_2 > 4$$

The F distribution depends on its degrees of freedom and is characterized by a positive skew. Figure 6.10 illustrates three different F density curves.

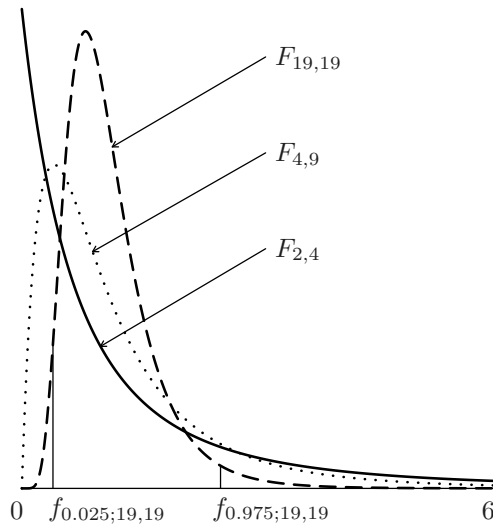


FIGURE 6.10: Illustrations of the pdfs of $F_{2,4}$ (solid line), $F_{4,9}$ (dotted line), and $F_{19,19}$ (dashed line) random variables

Theorem 6.5 If there are two random samples X_1, \dots, X_{n_X} and Y_1, \dots, Y_{n_Y} that are taken from independent normal populations where $X \sim N(\mu_X, \sigma_X)$ and $Y \sim N(\mu_Y, \sigma_Y)$, then the random variable

$$\frac{\frac{S_X^2}{\sigma_X^2}}{\frac{S_Y^2}{\sigma_Y^2}} \sim F_{n_X - 1, n_Y - 1}. \quad (6.28)$$

Proof: Since $\frac{S_X^2}{\sigma_X^2} \sim \frac{\chi_{n_X-1}^2}{n_X-1}$ and $\frac{S_Y^2}{\sigma_Y^2} \sim \frac{\chi_{n_Y-1}^2}{n_Y-1}$, by Theorem 6.3 on page 231, it follows that

$$\frac{\frac{S_X^2}{\sigma_X^2}}{\frac{S_Y^2}{\sigma_Y^2}} \sim F_{n_X-1, n_Y-1}.$$

To find the value $f_{\alpha; \nu_1, \nu_2}$, where $\mathbb{P}(F_{\nu_1, \nu_2} < f_{\alpha; \nu_1, \nu_2}) = \alpha$, with S, use the command `qf(p, df1, df2)`, where `p` is the area to the left (probability) in an F distribution with $\nu_1 = \text{df1}$ and $\nu_2 = \text{df2}$.

Example 6.26 Find the constants c and d such that $\mathbb{P}(F_{5,10} < c) = 0.95$ and $\mathbb{P}(F_{5,10} < d) = 0.05$.

Solution: Using the S commands `qf(0.95, 5, 10)` and `qf(0.05, 5, 10)` returns the values 3.325835 and 0.2111904, respectively. ■

Example 6.27 Use S to find the values associated with the points $f_{0.025; 19, 19}$ and $f_{0.975; 19, 19}$ depicted in Figure 6.10 on the previous page.

Solution: The answers using S are

```
> qf(.975, 19, 19)
[1] 2.526451
> qf(.025, 19, 19)
[1] 0.3958122
```

Note that a relationship exists between the t - and F distributions. Namely, $t_\nu^2 = F_{1, \nu}$, and the relationship between the values in both distributions is

$$t_{1-\alpha/2; \nu}^2 = f_{1-\alpha; 1, \nu}. \quad (6.29)$$

For example, $t_{0.975; 5}^2 = 2.571^2 = 6.61 = F_{0.95; 1, 5}$.

6.7 Problems

- How many ways can a host randomly choose 8 people out of 90 in the audience to participate in a TV game show?
- Let X be a t_5 .
 - Find $\mathbb{P}(X < 3)$.
 - Calculate $\mathbb{P}(2 < X < 3)$.
 - Find a so that $\mathbb{P}(X < a) = 0.05$.
- If $(1 - 2t)^{-5}$, $t < \frac{1}{2}$, is the **mgf** of a random variable X , find $\mathbb{P}(X < 15.99)$.
- If $X \sim \chi_{10}^2$, find the constants a and b so that $\mathbb{P}(a < X < b) = 0.90$ and $\mathbb{P}(X < a) = 0.05$.
- Let X be a χ_{10}^2 . Calculate $\mathbb{P}(X < 8)$ and $\mathbb{P}(X > 6)$. Calculate a so that $\mathbb{P}(X < a) = .05$. What are the population mean and population variance of X ?
- Let X be distributed as an $F_{2,5}$. Calculate $\mathbb{P}(X < 1)$ and the median of X . Calculate a so that $\mathbb{P}(X < a) = 0.10$. What are the population mean and population variance of X ?
- Assume a population with 5 elements:

$$X_1 = 0, \quad X_2 = 1, \quad X_3 = 2, \quad X_4 = 3, \quad X_5 = 4.$$

- Calculate \bar{X} and σ^2 .
 - Calculate the sampling distribution of the mean for random samples of size 3 taken without replacement. Verify that the mean of \bar{X} is 2 and that the variance of \bar{X} is $\sigma^2/6$.
 - Calculate the sampling distribution of \bar{X} for random samples of size 3 taken with replacement. Verify that the mean of \bar{X} is 2 and that the variance of \bar{X} is σ^2/n .
- A population has the following elements: 2, 5, 8, 12, 13.
 - Enumerate all the samples of size 2 that can be drawn with and without replacement.
 - Calculate the mean of the population.
 - Calculate the variance of the population.
 - Calculate the standard deviation of the population.
 - Calculate the mean of the sample mean, $E[\bar{X}]$.
 - Calculate the variance of the sampled mean, $Var(\bar{X})$.
 - Calculate the standard deviation of the sample mean.
 - Calculate the mean of the sample variance, $E[S^2]$.
 - Is the variance of \bar{X} larger when sampling with or without replacement? Explain your answer.
 - Determine whether the following expressions are statistics or not:
 - $\sum_{i=1}^n X_i$

- (b) $\sum_{i=1}^n X_i - \bar{X}$
 (c) $\bar{X} - \sigma$
 (d) $X_1 + X_2/6$
10. Use the data frame `wheatUSA2004` from the `PASWR` package; draw all samples of sizes 2, 3, and 4; and calculate the mean of the means. What size provides the best approximation to the population mean? What is the variance of these means?
11. Given a random sample of size 6 from $N(0, \sigma)$, calculate
- (a) $\mathbb{P}\left(\frac{\bar{X}}{S} > 2\right)$ and
 (b) $\mathbb{P}\left(\left|\frac{\bar{X}}{S_u}\right| \leq 4\right)$.
12. Constant velocity joints (CV joints) allow a rotating shaft to transmit power through a variable angle, at constant rotational speed, without an appreciable increase in friction or play. An after-market company produces CV joints. To optimize energy transfer, the drive shaft must be very precise. The company has two different branches that produce CV joints where the variability of the drive shaft is known to be 2 mm. A sample of $n_1 = 10$ is drawn from the first branch, and a sample of $n_2 = 15$ is drawn from the second branch. Suppose that the diameter follows a normal distribution. What is the probability that the drive shafts coming from the first branch will have greater variability than those of the second branch?
13. Given a population $N(\mu, \sigma)$ with unknown mean and variance, a sample of size 11 is drawn and the sample variance S^2 is calculated. Calculate the probability $\mathbb{P}(0.5 < S^2/\sigma^2 < 1.2)$.
14. Simulate 20,000 random samples of sizes 30, 100, 300, and 500 from an exponential distribution with a mean of $1/5$. Estimate the density of the sampling distribution with the function `density()`. Superimpose a theoretical normal density with appropriate mean and standard deviation. What sample size is needed to get an estimated density close to a normal density?
15. The plastic tubes produced by company X for the irrigation system used in golf courses have a length of 1.5 meters and a standard deviation of 0.1 meter. The plastic tubes produced by company Y have a length of 1 meter and a standard deviation of 0.09 meter. Suppose that both tube lengths follow normal distributions.
- (a) Calculate the probability that a random sample of 15 tubes from company X has a mean length at least 0.45 meter greater than the mean length of a random sample of size 20 from company Y .
- (b) Suppose that the population variances are unknown but equal, $S_x = 0.1$, and $S_y = 0.09$. Calculate the probability that a random sample of 15 plastic tubes from company X has a mean length at least 0.45 meter greater than the mean length of a random sample of 20 plastic tubes from company Y .
16. Plot the density function of an $F_{4,6}$ random variable. Find the area to the left of $x = 3$ and shade this region in the original plot.
17. Let X_1, X_2, X_3, X_4 be a random sample from a $N(0, \sigma)$. Calculate the distribution of $\frac{(X_1 - X_2)^2}{(X_3 + X_4)^2}$.

18. Let $X_1, X_2, X_3, X_4, X_5, X_6$ be a random sample drawn from a $N(0, \sigma^2)$ population. Find the values of c so that the statistic $\frac{cX_1 + X_2 + X_3}{\sqrt{X_4^2 + X_5^2 + X_6^2}}$ follows a t_3 -distribution.
19. Consider a random sample of size n from an exponential distribution with parameter λ . Use moment generating functions to show that the sample mean follows a $\Gamma(n, \lambda n)$. Graph the theoretical sampling distribution of \bar{X} when sampling from an $Exp(\lambda = 1)$ for $n = 30, 100, 300$, and 500 . Superimpose an appropriate normal density for each $\Gamma(n, \lambda n)$. At what sample size do the sampling distribution and superimposed density virtually coincide?
20. Set the seed equal to 10, and simulate 20,000 random samples of size $n_x = 65$ from a $N(4, \sigma_x = \sqrt{2})$, 20,000 random samples of size $n_y = 90$ from a $N(5, \sigma_y = \sqrt{3})$ and verify that the simulated statistic $\frac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2}$ follows an $F_{64,89}$ distribution.
21. Set the seed equal to 95, and simulate $m = 20,000$ random samples of size $n = 1000$ from a *Bernoulli*($\pi = 0.4$). Verify that the sample proportion follows an approximate normal distribution with a mean approximately equal to 0.4 and a standard deviation approximately equal to 0.01549.
22. A communication system consists of n components, where the probability that each component works is π . The system will work if at least half of its components work. For what values of π will a system consisting of 5 components have a greater probability of working than a system consisting of 3 components? Plot the probability each system ($n = 5$ and $n = 3$) works for values of π from 0 to 1 in increments of 0.01.
23. Given $X \sim N(0, \sigma = 1)$, $Y \sim N(2, \sigma = 2)$, and $Z \sim N(4, \sigma = 3)$, what is the distribution of $W = X + Y + Z$? Set the seed equal to 368 and simulate 1000 samples, each of size 1 for X , Y , and Z . Add the values in the three vectors to obtain W 's empirical distribution. Create a density histogram of the simulated values of W and superimpose the theoretical density of W .
24. Set the seed equal to 48, and simulate a χ_3^2 distribution by summing the squares of three simulated standard normal random variables, each having length 20,000. Create a density histogram of the simulated χ_3^2 random variable. Superimpose the theoretical χ_3^2 density over the histogram.
25. Verify empirically that

$$\frac{N(0, 1)}{(\frac{1}{5}\chi_5^2)^{\frac{1}{2}}} \sim t_5$$

by setting the seed equal to 36 and generating a sample of size 1000 from a $N(0, 1)$ distribution. Generate another sample of size 1000 from a χ_5^2 distribution. Perform the appropriate arithmetic to arrive at the simulated sampling distribution. Create a density histogram of the results and superimpose a theoretical t_5 density.

26. A farmer is interested in knowing the mean weight of his chickens when they leave the farm. Suppose that the standard deviation of the chickens' weight is 500 grams.
- (a) What is the minimum number of chickens needed to ensure the a standard deviation of the mean is no more than 100 grams with a confidence level of 0.95?
- (b) If the farm has three coops and the mean chicken weight in each coop is 1.8, 1.9, and 2 kg, respectively, calculate the probability that a random sample of 50 chickens with an average weight larger than 1.975 kg comes from the first coop. Assume the weight of the chickens follows a normal distribution.

27. Find the required sample size (n) to estimate the proportion of students spending more than €10 a week on entertainment with a 95% confidence interval so that the margin of error is no more than 0.02.
28. 15.3% of the Spanish Internet domain names are “.org.” If a sample of 2000 Spanish domain names is taken,
 - (a) Calculate the exact probability that at least 200 domain names will be “.org.”.
 - (b) Compute an approximate answer that at least 200 domain names will be “.org.” with a normal approximation.
29. Set the seed equal to 86, and simulate $m_1 = 20,000$ samples of size $n_1 = 1000$ from a $Bin(n_1, \pi = 0.3)$ and $m_2 = 20,000$ samples of size $n_2 = 1100$ from a $Bin(n_2, \pi = 0.7)$. Verify that the difference of sampling proportions follows a normal distribution.
30. Given a random sample of size n from an exponential distribution with parameter λ , prove that the sample mean follows a $\Gamma(n, \lambda n)$. Set the seed equal to 679, and simulate $m = 1000$ random samples of size $n = 100$ from an $Exp(\lambda = 1)$, and check that the normal approximation of the mean is appropriate. Repeat this exercise with random samples of size $n = 3$, and verify that, in this case, $\Gamma(3, 3)$ is more appropriate to use than the normal distribution.

Chapter 7

Point Estimation

7.1 Introduction

Throughout this chapter, random samples drawn from a known distribution where the parameters that characterize the distribution are unknown will be of interest. To specify completely a probability distribution, whether it be discrete or continuous, the distribution's parameters must be specified. For example, a random variable may follow a normal distribution; however, if both the mean and the standard deviation of the normal distribution are not known, the distribution at hand cannot be completely specified. In a similar fashion, a Poisson random variable requires knowledge of the parameter λ to specify completely that distribution. In general, the **pdf** of a random variable X is $f(x | \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the vector of parameters that characterize the **pdf**. The vector of parameters $\boldsymbol{\theta}$ is defined in a parameter space denoted Θ . For each value of $\boldsymbol{\theta} \in \Theta$, there is a different **pdf**. To obtain possible values for the vector of parameters, a random sample from the population of interest is taken and statistics called **estimators** are constructed. The values of the estimators are called **point estimates**. For example, \bar{X} may be used as a point estimator for μ , in which case \bar{x} is a point estimate of μ .

Since estimators are statistics or functions of random variables, they themselves are random variables. Studying the sampling distributions of estimators as well as their statistical properties such as mean square error, bias or unbiasedness, efficiency, consistency, and robustness, all of which will be defined in this chapter, will give guidelines about which estimators to employ.

7.2 Properties of Point Estimators

7.2.1 Mean Square Error

The goodness of an estimator is related to how close its estimates are to the true parameter. The difference between an estimator T for an unknown parameter θ and the parameter θ itself is called the error. Since this quantity can be either positive or negative, it is common to square the error so that various estimators T_1, T_2, \dots , can be compared using a non-negative measure of error. To that end, the **mean square error** of an estimator, denoted $MSE[T]$, is defined as $MSE[T] = E[(T - \theta)^2]$. Estimators with small MSE s will have a distribution such that the values in the distribution will be close to the true parameter. In fact, the MSE consists of two non-negative components, the variance of the

estimator T and the squared bias of the estimator T , where bias is defined as $E[T] - \theta$ since

$$\begin{aligned}
 MSE[T] &= E[(T - E[T] + E[T] - \theta)^2] \\
 &= E[T - E[T]]^2 + E[(E[T] - \theta)^2] + 2E[(T - E[T])(E[T] - \theta)] \\
 &= Var[T] + (E[T] - \theta)^2 + 2(E[T] - E[T])(E[T] - \theta) \\
 &= Var[T] + (E[T] - \theta)^2 \\
 &= Var[T] + (Bias[T])^2.
 \end{aligned}
 \tag{7.1}$$

The concepts of variance and bias are illustrated in Figure 7.1, which depicts the shot patterns for four marksmen on their respective targets. When the marksman's weapon is properly sighted, the center of the target represents θ .

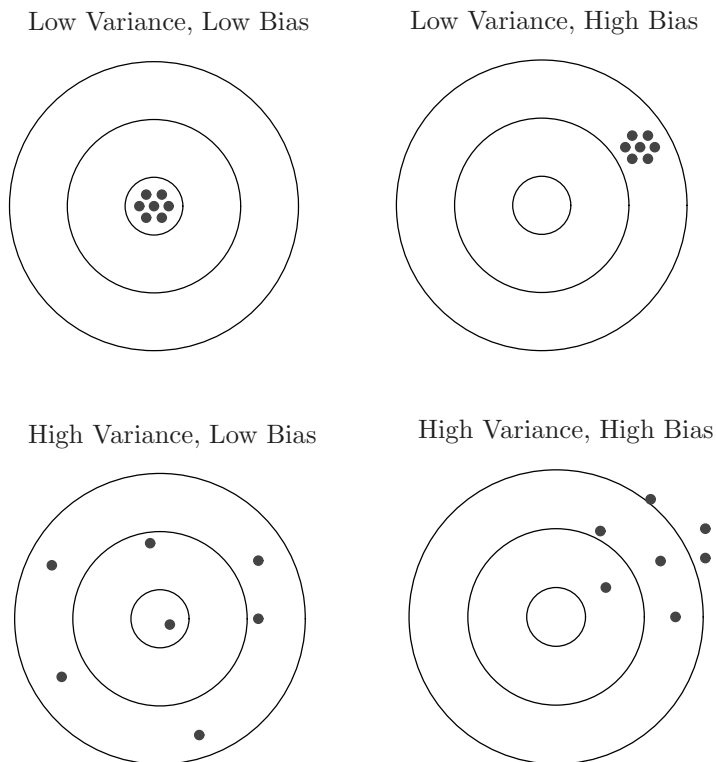


FIGURE 7.1: Visual representations of variance and bias

It seems logical to think that the most desirable estimators are those that minimize the MSE . However, estimators that minimize the MSE for all possible values of θ do not always exist. In other words, an estimator may have the minimum MSE for some values of θ and not others.

7.2.2 Unbiased Estimators

Since estimators are random variables, the point estimates they return will vary from sample to sample. However, one would like some assurance that the chosen estimator is returning a value close to the unknown parameter. Estimators whose expected values are equal to the parameters they are estimating are **unbiased**. That is, when $E[T] = \theta$, T is an **unbiased** estimator of θ . When an estimator is unbiased, its MSE is equal to its variance, that is, $MSE[T] = Var[T]$. On the other hand, when $E[T] \neq \theta$, the estimator is biased.

Example 7.1 Show that the sample mean and the sample variance are unbiased estimators of the population mean and the population variance, respectively.

Solution: To show that S^2 is an unbiased estimator of σ^2 , use the fact that

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n (X_i - \mu)^2 - n(\mu - \bar{X})^2$$

from (6.14) on page 225:

$$\begin{aligned} E[\bar{X}] &= E\left[\sum_{i=1}^n \frac{X_i}{n}\right] = \frac{n\mu}{n} = \mu \\ E[S^2] &= E\left[\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}\right] = E\left[\frac{\sum_{i=1}^n (X_i - \mu)^2 - n(\mu - \bar{X})^2}{n-1}\right] \\ &= \frac{1}{n-1} \left[\sum_{i=1}^n E[(X_i - \mu)^2] - nE[(\bar{X} - \mu)^2] \right] \\ &= \frac{1}{n-1} \left[n\sigma^2 - n\frac{\sigma^2}{n} \right] = \sigma^2 \end{aligned}$$

Example 7.2 Suppose $X \sim Pois(\lambda)$, where λ is unknown. Show

- \bar{X} is an unbiased estimator of λ .
- $2\bar{X}$ is an unbiased estimator of 2λ .
- \bar{X}^2 is a biased estimator of λ^2 .

Solution: To solve the problems, keep in mind that if $X \sim Pois(\lambda)$, $E[X] = \lambda$ and $Var[X] = \lambda$.

- Since $E[\bar{X}] = E\left[\sum_{i=1}^n \frac{X_i}{n}\right] = \sum_{i=1}^n \frac{E[X_i]}{n} = \frac{n\lambda}{n} = \lambda$, it follows that \bar{X} is an unbiased estimator of λ .
- Since $E[2\bar{X}] = 2E[\bar{X}] = 2\lambda$, it follows that $2\bar{X}$ is an unbiased estimator of 2λ .
- Since $E[\bar{X}^2] = Var[\bar{X}] + \mu_{\bar{X}}^2 = \frac{\lambda}{n} + \lambda^2$, it follows that \bar{X}^2 is a biased estimator of λ^2 . However, \bar{X}^2 is an asymptotically unbiased estimator of λ^2 . That is, as n tends to infinity, the estimator becomes unbiased. ■

Example 7.3 Suppose $\{X_1, X_2, \dots, X_n\}$ is a random sample from a $N(\mu, \sigma)$ distribution. Show that S is a biased estimator of σ .

Solution: Recall that $\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$. Let $X = \frac{(n-1)S^2}{\sigma^2}$ and take the square root and the expected value of both sides:

$$E[\sqrt{X}] = E\left[\frac{\sqrt{n-1}}{\sigma} \cdot S\right].$$

Since $X \sim \chi_{n-1}^2$, the expected value of \sqrt{X} is $\int_{-\infty}^{\infty} \sqrt{x}f(x)dx$, where $f(x)$ is the **pdf** of a chi-square random variable:

$$\begin{aligned} E[\sqrt{X}] &= \int_0^{\infty} \sqrt{x} \frac{1}{\Gamma\left(\frac{n-1}{2}\right) 2^{\frac{n-1}{2}}} x^{\frac{n-1}{2}-1} e^{-\frac{x}{2}} dx \\ &= \frac{1}{\Gamma\left(\frac{n-1}{2}\right) 2^{\frac{n-1}{2}}} \int_0^{\infty} x^{\frac{n-1}{2}-1+\frac{1}{2}} e^{-\frac{x}{2}} dx \\ &= \frac{1}{\Gamma\left(\frac{n-1}{2}\right) 2^{\frac{n-1}{2}}} \int_0^{\infty} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} dx \end{aligned} \quad (7.2)$$

Next, use the change of variable $x/2 = t$ where $dx = 2dt$ in an attempt to force the right-hand side of (7.2) to look like a gamma function. Specifically, recall that $\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx$ for $\alpha > 0$:

$$\begin{aligned} E[\sqrt{X}] &= \frac{1}{\Gamma\left(\frac{n-1}{2}\right) 2^{\frac{n-1}{2}}} \int_0^{\infty} (2t)^{\frac{n}{2}-1} e^{-t} 2 dt \\ &= \frac{2^{\frac{n}{2}}}{\Gamma\left(\frac{n-1}{2}\right) 2^{\frac{n-1}{2}}} \int_0^{\infty} t^{\frac{n}{2}-1} e^{-t} dt = \frac{\sqrt{2}\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \end{aligned}$$

Since

$$E[\sqrt{X}] = E\left[\frac{\sqrt{n-1}}{\sigma} S\right] = \frac{\sqrt{2}\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)},$$

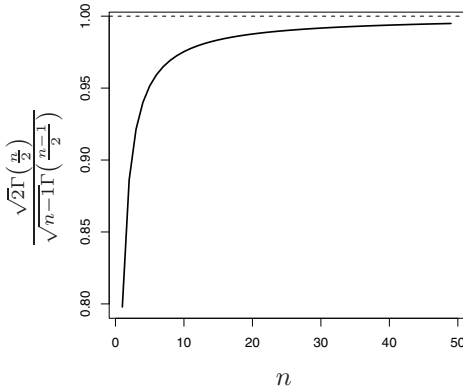
it follows that

$$E[S] = \sigma \frac{\sqrt{2}\Gamma\left(\frac{n}{2}\right)}{\sqrt{n-1}\Gamma\left(\frac{n-1}{2}\right)} \neq \sigma \quad (7.3)$$

Therefore, S is a biased estimator of σ . ■

Example 7.4 Numerically evaluate and graph the coefficient $\frac{\sqrt{2}\Gamma\left(\frac{n}{2}\right)}{\sqrt{n-1}\Gamma\left(\frac{n-1}{2}\right)}$ that multiplies σ on the right-hand side of (7.3) for values of n from 2 to 50.

Solution: The following S code creates a graph similar to the one depicted to the left of the code. Note that the coefficient $\frac{\sqrt{2}\Gamma\left(\frac{n}{2}\right)}{\sqrt{n-1}\Gamma\left(\frac{n-1}{2}\right)}$ is virtually 1 for values of $n \geq 20$, so that S is a reasonable, though biased, estimator of σ for $n \geq 20$. Note that in the following code, no true `coeff` value is assigned to `coeff[1]` because $n-1$ would then be zero. Therefore, when `coeff` is plotted, `coeff[1]` is removed with `coeff[-1]`.



```

> m <- 50
> coeff <- array(0, m)
> for (n in 2:m)
+ { coeff[n] <- (sqrt(2/(n-1))
+ *gamma(n/2))/gamma((n-1)/2)}
> plot(coeff[-1], type="l",
+ xlab="n", ylab="coef", lwd=2)
> abline(h=1, lty=2)
> coeff[20]
[1] 0.9869343

```

A more compact solution using R is

```

> curve(sqrt(2/(x-1))*gamma(x/2)/gamma((x-1)/2), 2, 50)
> abline(h=1, lty=2)

```

7.2.3 Efficiency

A desirable property of a good estimator is not only to be unbiased, but also to have a small variance, which translates into a small *MSE* for estimators, regardless of whether they are biased or unbiased. One way to compare the *MSEs* of two estimators is by using **relative efficiency**. Given two estimators T_1 and T_2 , the efficiency of T_1 relative to T_2 , written $eff(T_1, T_2)$, is

$$eff(T_1, T_2) = \frac{MSE[T_2]}{MSE[T_1]}. \quad (7.4)$$

When the estimators in (7.4) are unbiased, the efficiency of T_1 relative to T_2 is simply the ratio of estimators variances, written

$$eff(T_1, T_2) = \frac{Var[T_2]}{Var[T_1]}.$$

The estimator T_1 is more efficient than the estimator T_2 if, for any sample size, $MSE[T_1] \leq MSE[T_2]$, which then implies that $eff(T_1, T_2) \geq 1$. When the estimators are unbiased, the estimator T_1 is more efficient than the estimator T_2 if, for any sample size, $Var[T_1] \leq Var[T_2]$, which also implies that $eff(T_1, T_2) \geq 1$. If a choice is to be made among a small number of unbiased estimators, simply compute the variance of all of the estimators and select the estimator with minimum variance. However, if the estimator that has the smallest variance among all possible unbiased estimators must be chosen, an infinite number of variances would need to be calculated. Clearly, this is not a viable solution.

Thankfully, it can be shown that if $T = \hat{\theta}$ is an unbiased estimator of θ and a random sample of size n , X_1, X_2, \dots, X_n , has **pdf** $f(x|\theta)$, then the variance of the unbiased estimator, $\hat{\theta}$, must satisfy the inequality

$$Var[\hat{\theta}] \geq \frac{1}{n \cdot E \left[\left(\frac{\partial \ln f(X|\theta)}{\partial \theta} \right)^2 \right]}, \quad (7.5)$$

where $f(X|\theta)$ is the density function of the distribution of interest evaluated at the random variable X . In the discrete case, $p(X|\theta)$ is used instead of $f(X|\theta)$. In general, the probability

distributions of both discrete and continuous distributions are referred to using the notation $f(x)$. The inequality in (7.5) is known as the **Cramér-Rao inequality**, and the quantity on the right-hand side of the equation is known as the Cramér-Rao lower bound (CRLB).

DEFINITION 7.1: If $\hat{\theta}$ is an unbiased estimator of θ and

$$\text{Var}[\hat{\theta}] = \frac{1}{n \cdot E \left[\left(\frac{\partial \ln f(X|\theta)}{\partial \theta} \right)^2 \right]} \quad (7.6)$$

then $\hat{\theta}$ is a **minimum variance unbiased** estimator of θ .

Not all parameters have unbiased estimators whose variance equals the CRLB. However, when the variance of an unbiased estimator equals the CRLB, the estimator is **efficient** or **minimum variance**. The quantity in the denominator of (7.6) is known as the **Fisher information** about θ that is supplied by the sample. That is, the smaller the variance of the estimator, the greater the information.

Example 7.5 Show that \bar{X} is a minimum variance unbiased estimator of the mean λ of a Poisson population.

Solution: If $X \sim \text{Pois}(\lambda)$, then, according to (4.5), $E[X] = \lambda$, $\text{Var}[X] = \lambda$, and the **pdf** of X is

$$\mathbb{P}(X = x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}. \quad (7.7)$$

Since $E[\bar{X}] = \sum_{i=1}^n \frac{E[X_i]}{n} = \frac{n\lambda}{n} = \lambda$, it follows that \bar{X} is an unbiased estimator of λ , with variance $\frac{\lambda}{n}$ because the $\text{Var}[\bar{X}] = \text{Var}[\sum_{i=1}^n \frac{X_i}{n}] = \frac{1}{n^2} \sum_{i=1}^n \text{Var}[X_i] = \frac{n\lambda}{n^2} = \frac{\lambda}{n}$. Consequently, if the CRLB equals $\frac{\lambda}{n}$, \bar{X} is a minimum variance unbiased estimator of λ according to Definition 7.1. By taking the natural logarithm of (7.7),

$$\ln \mathbb{P}(x|\lambda) = x \ln(\lambda) - \lambda - \ln(x!). \quad (7.8)$$

Taking the derivative of (7.8) with respect to λ gives

$$\frac{\partial \ln \mathbb{P}(x|\lambda)}{\partial \lambda} = \frac{x}{\lambda} - 1 = \frac{x - \lambda}{\lambda}.$$

Hence

$$E \left[\left(\frac{\partial \ln \mathbb{P}(X|\lambda)}{\partial \lambda} \right)^2 \right] = E \left[\left(\frac{X - \lambda}{\lambda} \right)^2 \right] = \frac{E[(X - \lambda)^2]}{\lambda^2} = \frac{\text{Var}[X]}{\lambda^2}.$$

Therefore,

$$E \left[\left(\frac{\partial \ln \mathbb{P}(X|\lambda)}{\partial \lambda} \right)^2 \right] = \frac{\text{Var}[X]}{\lambda^2} = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda},$$

and the CRLB is

$$\frac{1}{n \cdot E \left[\left(\frac{\partial \ln f(X|\lambda)}{\partial \lambda} \right)^2 \right]} = \frac{\lambda}{n}.$$

Consequently, since \bar{X} is unbiased and $\text{Var}[\bar{X}] = \frac{\lambda}{n}$, it follows that \bar{X} is a minimum variance unbiased estimator of λ . ■

Example 7.6 Show that \bar{X} is a minimum variance unbiased estimator of the mean θ of an exponential population.

Solution: If $X \sim \text{Exp}(\frac{1}{\theta})$, then, according to (4.12), when using the substitution $\theta = \frac{1}{\lambda}$, $E[X] = \theta$, $\text{Var}[X] = \theta^2$, and the **pdf** of X is

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-x/\theta} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}. \quad (7.9)$$

Since $E[\bar{X}] = \sum_{i=1}^n \frac{E[X_i]}{n} = \frac{n\theta}{n} = \theta$, it follows that \bar{X} is an unbiased estimator of θ , with variance $\frac{\theta^2}{n}$ since $\text{Var}[\bar{X}] = \text{Var}[\sum_{i=1}^n \frac{X_i}{n}] = \frac{1}{n^2} \sum_{i=1}^n \text{Var}[X_i] = \frac{n\theta^2}{n^2} = \frac{\theta^2}{n}$. Consequently, if the CRLB equals $\frac{\theta^2}{n}$, \bar{X} is a minimum variance unbiased estimator of θ according to Definition 7.1 on the facing page. By taking the natural logarithm of (7.9),

$$\ln f(x|\theta) = -\ln(\theta) - \frac{x}{\theta}. \quad (7.10)$$

Taking the derivative of (7.10) with respect to θ gives

$$\frac{\partial \ln f(x|\theta)}{\partial \theta} = -\frac{1}{\theta} + \frac{x}{\theta^2} = \frac{x - \theta}{\theta^2}.$$

Hence

$$E \left[\left(\frac{\partial \ln f(X|\theta)}{\partial \theta} \right)^2 \right] = E \left[\left(\frac{X - \theta}{\theta^2} \right)^2 \right] = \frac{E[(X - \theta)^2]}{\theta^4} = \frac{\text{Var}[X]}{\theta^4}.$$

Therefore,

$$E \left[\left(\frac{\partial \ln f(X|\lambda)}{\partial \theta} \right)^2 \right] = \frac{\text{Var}[X]}{\theta^4} = \frac{\theta^2}{\theta^4} = \frac{1}{\theta^2},$$

and the CRLB is

$$\frac{1}{n \cdot E \left[\left(\frac{\partial \ln f(X|\theta)}{\partial \theta} \right)^2 \right]} = \frac{\theta^2}{n}.$$

Consequently, since \bar{X} is unbiased and $\text{Var}[\bar{X}] = \frac{\theta^2}{n}$, it follows that \bar{X} is a minimum variance unbiased estimator of θ . ■

Example 7.7 ▷ *Comparing Estimators: Blue Jean Length* ◁ Suppose the true manufactured length of new 32L blue jeans follows a normal distribution with unknown μ and $\sigma = 0.5$ inch. It is known that 32L blue jeans sold in stores have a length of at least 31 inches. If a random sample of size $n = 3$ of 32L blue jeans is taken to estimate μ , which of the estimators $\hat{\mu}_1$ or $\hat{\mu}_2$ is better in terms of bias, variance, and relative efficiency where $\hat{\mu}_1 = 0.33 \cdot (X_1 + X_2 + X_3)$ and $\hat{\mu}_2 = 0.50 \cdot (X_1 + X_2)$?

Solution: Since

$$\begin{aligned} E[\hat{\mu}_1] &= 0.33 \cdot E[X_1 + X_2 + X_3] = 0.33 \cdot (E[X_1] + E[X_2] + E[X_3]) \\ &= 0.33(\mu + \mu + \mu) = 0.99\mu, \end{aligned}$$

it follows that $\hat{\mu}_1$ is a biased estimator of μ with bias $0.99\mu - \mu = -0.01\mu$. On the other hand,

$$E[\hat{\mu}_2] = 0.50 \cdot E[X_1 + X_2] = 0.50 \cdot (E[X_1] + E[X_2]) = 0.50 \cdot (\mu + \mu) = \mu,$$

which makes $\hat{\mu}_2$ an unbiased estimator of μ . The variances of $\hat{\mu}_1$ and $\hat{\mu}_2$ are

$$\begin{aligned} \text{Var}[\hat{\mu}_1] &= \text{Var}[0.33 \cdot (X_1 + X_2 + X_3)] \\ &= 0.33^2 \cdot (\text{Var}[X_1] + \text{Var}[X_2] + \text{Var}[X_3]) \\ &= 0.33^2 \cdot (0.25 + 0.25 + 0.25) = 0.081675, \text{ and} \\ \text{Var}[\hat{\mu}_2] &= \text{Var}[0.50 \cdot (X_1 + X_2)] = 0.50^2 \cdot (\text{Var}[X_1] + \text{Var}[X_2]) \\ &= 0.25 \cdot (0.25 + 0.25) = 0.125, \text{ respectively.} \end{aligned}$$

Before looking at the relative efficiency of $\hat{\mu}_1$ to $\hat{\mu}_2$, compute the *MSE* for each estimator using the fact that $MSE = \text{Variance} + \text{Bias}^2$:

$$\begin{aligned} MSE[\hat{\mu}_1] &= 0.081675 + (0.01\mu)^2 = 0.081675 + 0.0001\mu^2 \\ MSE[\hat{\mu}_2] &= 0.125 + 0^2 = 0.125 \end{aligned}$$

Since

$$eff(\hat{\mu}_1, \hat{\mu}_2) = \frac{MSE(\hat{\mu}_2)}{MSE(\hat{\mu}_1)} = \frac{0.125}{0.081675 + 0.0001\mu^2} < 1 \text{ for all } |\mu| > 20.82,$$

conclude that $\hat{\mu}_2$ is both more efficient and has a smaller *MSE* than does $\hat{\mu}_1$, since it is known that $\mu \geq 31$ inches according to the problem. See Figure 7.2 for a graphical representation of the distributions of $\hat{\mu}_1$ and $\hat{\mu}_2$.

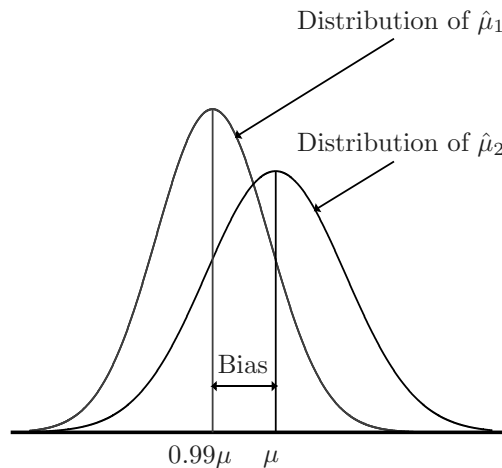


FIGURE 7.2: Graphical representations for the sampling distributions of $\hat{\mu}_1$ and $\hat{\mu}_2$

■

7.2.4 Consistent Estimators

The next property of estimators that is considered is **consistency**. Consistency is a property of a sequence of estimators rather than a single estimator. However, it is rather common to refer to an estimator as being consistent. A sequence of estimators means that the same estimation procedure is carried out for each sample of size n . If T is an estimator

of θ and X_1, X_2, \dots are observed according to a distribution $f(x|\theta)$, a sequence of estimators T_1, T_2, \dots, T_n can be constructed by performing the same estimation procedure for samples of sizes $1, 2, \dots, n$, respectively. In other words, the sequence is

$$T_1 = t(X_1), T_2 = t(X_1, X_2), \dots, T_n = t(X_1, X_2, \dots, X_n).$$

A sequence of estimators T_n (defined for all n) is a **consistent** estimator of the parameter θ for every $\theta \in \Theta$ if

$$\lim_{n \rightarrow \infty} \mathbb{P}(|T_n - \theta| \geq \epsilon) = 0, \text{ for all } \epsilon > 0. \quad (7.11)$$

An equivalent statement of (7.11) is that a sequence of estimators T_n (defined for all n) is a **consistent** estimator of the parameter θ for every $\theta \in \Theta$ if

$$\lim_{n \rightarrow \infty} \mathbb{P}(|T_n - \theta| < \epsilon) = 1, \text{ for all } \epsilon > 0. \quad (7.12)$$

Both definitions (7.11) and (7.12) state that a consistent sequence of estimators **converges in probability** to the parameter θ , where θ is the parameter the consistent sequence of estimators is estimating. In practical terms, this implies that the variance of a consistent estimator decreases as n increases and that the expected value of T_n tends to θ as n increases. Further, given a consistent sequence of estimators, say T_n , Chebyshev's inequality (3.17) guarantees that

$$\mathbb{P}(|T_n - \theta| \geq \epsilon) = \mathbb{P}(|T_n - \theta|^2 \geq \epsilon^2) \leq \frac{E[(T_n - \theta)^2]}{\epsilon^2},$$

for every $\theta \in \Theta$. Since $E_\theta [(T_n - \theta)^2]$ can be expressed as

$$E_\theta [(T_n - \theta)^2] = \text{Var}[T_n] + (\text{Bias}[T_n])^2,$$

if

$$\lim_{n \rightarrow \infty} \text{Var}[T_n] = 0 \quad \text{and} \quad \lim_{n \rightarrow \infty} (\text{Bias}[T_n])^2 = 0, \quad (7.13)$$

then T_n is a consistent sequence of estimators of θ . Whenever the conditions in (7.13) are true, T_n converges in *MSE* to the true value of θ . The conditions in (7.13) are sufficient but not necessary conditions for a sequence of estimators to be consistent.

Example 7.8 Let $\{X_1, X_2, \dots, X_n\}$ be a random sample of size n from a distribution with mean μ and variance σ^2 . Show that \bar{X}_n is a consistent estimator of μ .

Solution: For \bar{X}_n to be a consistent estimator of μ , it must be shown that

$$\lim_{n \rightarrow \infty} \mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) = 0 \text{ for all } \epsilon > 0.$$

Using version (c) of Chebyshev's inequality and the fact that $E[\bar{X}_n] = \mu$ and $\text{Var}[\bar{X}_n] = \sigma^2/n$,

$$\mathbb{P}(|\bar{X}_n - \mu| \geq k\sigma/\sqrt{n}) \leq \frac{1}{k^2}.$$

By setting $\epsilon = k\sigma/\sqrt{n}$, $k = \sqrt{n}\epsilon/\sigma$, so that

$$\frac{1}{k^2} = \frac{\sigma^2}{n\epsilon^2},$$

from which it follows that

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}. \quad (7.14)$$

Given that $\sigma^2 < \infty$ (finite), taking the limit as $n \rightarrow \infty$ on both sides of the \leq sign of (7.14) gives

$$\lim_{n \rightarrow \infty} \mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) = 0 \text{ for all } \epsilon.$$

Consequently, \bar{X}_n is a consistent estimator of μ . This is essentially the **weak law of large numbers** given in (3.18) of Section 3.4.7. ■

7.2.5 Robust Estimators

The idea of statistical **robustness** has received considerable attention in recent years. However, there is not a consensus on what defines a robust estimator. The essence of a **robust** estimator is an estimator whose sampling distribution is not seriously affected by violations of underlying assumptions. For example, when estimating the average useful life of an electronic component, one may think that an exponential distribution is being sampled when in fact a gamma or Weibull distribution is being sampled. If the estimation of the unknown parameter is not seriously affected by the fact that an incorrect distribution is being assumed, the estimator is robust. The concept of robustness has also been used to refer to the ability of a particular estimator to provide reasonable estimates when atypical observations are encountered in the sample. For example, if the largest value in a sample is made 1000 times larger, the sample median remains the same in both the sample with the original value and in the sample where the value is 1000 times larger than the largest value in the original sample. In this sense, the median is a robust estimator.

In particular, the median provides a robust measure of center whenever the underlying distribution is skewed. In a similar fashion, a robust measure of variability is the **median absolute deviation** (*MAD*). The *MAD* is defined as

$$MAD = \text{median}|x_i - \text{sample median}|. \quad (7.15)$$

When working with normal distributions, a robust estimator of σ is *MAD1*, where $MAD1 = \frac{1}{0.6745} MAD$. The value 0.6745 corresponds to the 75th percentile of a $N(0, 1)$ distribution ($z_{0.75} = 0.6745$). When working with S, the default value returned when working with the function `mad()` corresponds to the definition of *MAD1*. To compute the *MAD* as defined in (7.15), use the S option `constant=1` inside the `mad()` function.

Example 7.9 A botanist interested in studying the effects of a new herbicide on *trifolium repens* (white clover) measures and records the stem lengths in centimeters of ten specimens as 5.3, 2.8, 3.4, 7.2, 8.3, 1.7, 6.2, 9.3, 3.2, and 5.9. Compute the mean, median, standard deviation, and *MAD*. Suppose the botanist makes a field error and records an 83 instead of an 8.3. What effect will the recording error have on the computed quantities?

Solution: The stem measurements are entered without the recording error in the vector `stem1` (in increasing order) and the stem measurements with the recording error in the vector `stem2`. That is, `stem2` has an 83 rather than an 8.3.

```
> stem1 <- c(1.7, 2.8, 3.2, 3.4, 5.3, 5.9, 6.2, 7.2, 8.3, 9.3)
> stem2 <- c(1.7, 2.8, 3.2, 3.4, 5.3, 5.9, 6.2, 7.2, 83, 9.3)
> c(mean(stem1), sqrt(var(stem1)))
[1] 5.330000 2.516634
> c(mean(stem2), sqrt(var(stem2)))
[1] 12.80000 24.77185
> c(median(stem1), mad(stem1, constant = 1))
[1] 5.6 2.3
```

```

> c(median(stem2), mad(stem2, constant = 1))
[1] 5.6 2.3
> median(abs(stem1 - median(stem1)))
[1] 2.3
> median(abs(stem2 - median(stem2)))
[1] 2.3

```

Note that the mean and standard deviation of `stem1` (5.33, 2.52) are dramatically different from the mean and standard deviation of `stem2` (12.8, 24.77). However, the median and *MAD* (5.6, 2.3) are the same for the values in both `stem1` and `stem2`. What has been demonstrated is the robustness of the median and the *MAD* to outliers. ■

7.3 Point Estimation Techniques

Section 7.2 discussed several ways to measure the “goodness” of an estimator. In what follows, the framework for deriving estimators is provided. In general, these topics are intertwined. Specifically two methods are considered: the method of moments and the method of maximum likelihood. Before proceeding further, some notation is emphasized. Recall that capital letters are used to denote random variables. Specifically, the information in a random sample X_1, X_2, \dots, X_n is used to make inferences about the unknown θ . The observed values of the random sample are denoted x_1, x_2, \dots, x_n . Further, a random sample X_1, X_2, \dots, X_n is referred to with the boldface \mathbf{X} and the observed values in a random sample x_1, x_2, \dots, x_n with the boldface \mathbf{x} . The joint **pdf** of X_1, X_2, \dots, X_n is given by

$$\begin{aligned}
 f(\mathbf{x}|\theta) &= f(x_1, x_2, \dots, x_n|\theta) \\
 &= f(x_1|\theta) \times f(x_2|\theta) \times \dots \times f(x_n|\theta) = \prod_{i=1}^n f(x_i|\theta).
 \end{aligned}
 \tag{7.16}$$

7.3.1 Method of Moments Estimators

The idea behind the **method of moments** is to equate population moments about the origin to their corresponding sample moments, where the r^{th} **sample moment about the origin**, denoted m_r , is defined as

$$m_r = \frac{1}{n} \sum_{i=1}^n X_i^r,
 \tag{7.17}$$

and subsequently to solve for estimators of the unknown parameters. Recall that the r^{th} population moment about the origin of a random variable X , denoted α_r , was defined in (3.6) as $E[X^r]$. It follows that $\alpha_r = E[X^r] = \sum_{i=1}^{\infty} x_i^r \mathbb{P}(X = x_i)$ for discrete X , and that $\alpha_r = E[X^r] = \int_{-\infty}^{\infty} x^r f(x) dx$ for continuous X . Specifically, given a random sample X_1, X_2, \dots, X_n from a population with **pdf** $f(x|\theta_1, \theta_2, \dots, \theta_k)$, the method of moments estimators, denoted $\tilde{\theta}_i$ for $i = 1, \dots, k$, are found by equating the first k population moments about the origin to their corresponding sample moments and solving the resulting system

of simultaneous equations:

$$\left\{ \begin{array}{l} \alpha_1(\theta_1, \dots, \theta_k) = m_1 \\ \alpha_2(\theta_1, \dots, \theta_k) = m_2 \\ \quad \quad \quad \vdots \quad \quad \quad \vdots \\ \alpha_k(\theta_1, \dots, \theta_k) = m_k \end{array} \right. \quad (7.18)$$

The method of moments is an appealing technique for deriving estimators due to its simplicity and to the fact that method of moments estimators are consistent. In fact, the theoretical justification for equating the sample moments to the population moments is that, under certain conditions, it can be shown that the sample moments converge in probability to the population moments and that the sample moments about the origin are unbiased estimators of their corresponding population moments.

Example 7.10 Given a random sample of size n from a $Bin(1, \pi)$ population, find the method of moments estimator of π .

Solution: The first sample moment m_1 is \bar{X} , and the first population moment about zero for the binomial random variable is $\alpha_1 = E[X^1] = 1 \cdot \pi$. By equating the first population moment to the first sample moment,

$$\alpha_1(\pi) = \pi \stackrel{\text{set}}{=} \bar{X} = m_1,$$

which implies that the method of moments estimator for π is $\tilde{\pi} = \bar{X}$. ■

Example 7.11 Given a random sample of size m from a $Bin(n, \pi)$ population, find the method of moments estimator of π .

Solution: The first sample moment m_1 is \bar{X} , and the first population moment about zero for the binomial random variable is $\alpha_1 = E[X^1] = n \cdot \pi$. By equating the first population moment to the first sample moment,

$$\alpha_1(\pi) = n\pi \stackrel{\text{set}}{=} \bar{X} = m_1,$$

which implies that the method of moments estimator for π is $\tilde{\pi} = \frac{\bar{X}}{n}$. ■

Example 7.12 Given a random sample of size n from a $Pois(\lambda)$ population, find the method of moments estimator of λ .

Solution: The first sample moment m_1 is \bar{X} , and the first population moment about zero for a Poisson random variable is $\alpha_1 = E[X^1] = \lambda$. By equating the first population moment to the first sample moment,

$$\alpha_1(\pi) = \lambda \stackrel{\text{set}}{=} \bar{X} = m_1,$$

which implies that the method of moments estimator for λ is $\tilde{\lambda} = \bar{X}$. ■

Example 7.13 Given a random sample of size n from a $N(\mu, \sigma)$ population, find the method of moments estimators of μ and σ^2 .

Solution: The first and second sample moments m_1 and m_2 are \bar{X} and $\frac{1}{n} \sum_{i=1}^n X_i^2$, respectively. The first and second population moments about zero for a normal random variable are $\alpha_1 = E[X^1] = \mu$ and $\alpha_2 = E[X^2] = \sigma^2 + \mu^2$. By equating the first two population moments to the first two sample moments,

$$\begin{cases} \alpha_1(\mu, \sigma^2) = \mu \stackrel{\text{set}}{=} \bar{X} = m_1 \\ \alpha_2(\mu, \sigma^2) = \sigma^2 + \mu^2 \stackrel{\text{set}}{=} \frac{1}{n} \sum_{i=1}^n X_i^2 = m_2. \end{cases} \quad (7.19)$$

Solving the system of equations in (7.19) yields $\tilde{\mu} = \bar{X}$ and $\tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2 = S_u^2$ as the method of moments estimators for μ and σ^2 , respectively. ■

Example 7.14 Given a random sample of size n from a $Gamma(\alpha, \lambda)$ population, find the method of moments estimators of α and λ .

Solution: According to (4.16), $E[X] = \frac{\alpha}{\lambda}$, and $Var[X] = \frac{\alpha}{\lambda^2}$ for a random variable X that follows a gamma distribution. The first and second sample moments m_1 and m_2 are \bar{X} and $\frac{1}{n} \sum_{i=1}^n X_i^2$, respectively. The first and second population moments for a gamma random variable are

$$\alpha_1 = E[X^1] = \frac{\alpha}{\lambda},$$

and

$$\alpha_2 = E[X^2] = \sigma^2 + E[X]^2 = \frac{\alpha}{\lambda^2} + \frac{\alpha^2}{\lambda^2} = \frac{\alpha(1 + \alpha)}{\lambda^2},$$

respectively. By equating the first two population moments to the first two sample moments,

$$\begin{cases} \alpha_1(\alpha, \lambda) = \frac{\alpha}{\lambda} \stackrel{\text{set}}{=} \bar{X} = m_1 \\ \alpha_2(\alpha, \lambda) = \frac{\alpha(1 + \alpha)}{\lambda^2} \stackrel{\text{set}}{=} \frac{1}{n} \sum_{i=1}^n X_i^2 = m_2. \end{cases} \quad (7.20)$$

When it is recalled that $S_u^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$, the system of equations in (7.20) can be solved to obtain $\tilde{\alpha} = \frac{\bar{X}^2}{S_u^2}$ and $\tilde{\lambda} = \frac{\bar{X}}{S_u^2}$ as the method of moments estimators for α and λ , respectively. ■

7.3.2 Likelihood and Maximum Likelihood Estimators

When sampling from a population described by a **pdf** $f(x|\theta)$, knowledge of θ provides knowledge of the entire population. The idea behind maximum likelihood is to select the value for θ that makes the observed data most likely under the assumed probability model. When x_1, x_2, \dots, x_n are the observed values of a random variable X from a population with parameter θ , the notation $L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)$ will be used to indicate that the distribution depends on the parameter θ , and \mathbf{x} to indicate the distribution is dependent on the observed values from the sample. Once the sample values are observed, $L(\theta|\mathbf{x})$ can still be evaluated in a formal sense, although it no longer has a probability interpretation (in the discrete case) as does (7.16). $L(\theta|\mathbf{x})$ is the **likelihood function** of θ for \mathbf{x} and is denoted by

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta) = \prod_{i=1}^n f(x_i|\theta) = f(x_1|\theta) \times f(x_2|\theta) \times \cdots \times f(x_n|\theta). \quad (7.21)$$

The key difference between (7.16) and (7.21) is that the joint **pdf** given in (7.16) is a function of \mathbf{x} for a given θ and the likelihood function given in (7.21) is a function of θ for given \mathbf{x} .

The value of θ that maximizes $L(\theta|\mathbf{x})$ is called the **maximum likelihood estimate** (mle) of θ . Another way to think of the mle is the mode of the likelihood function. The maximum likelihood estimate is denoted as $\hat{\theta}(\mathbf{x})$, and the maximum likelihood estimator (MLE), a statistic, as $\hat{\theta}(\mathbf{X})$. In general, the likelihood function may be difficult to manipulate, and it is usually more convenient to work with the natural logarithm of $L(\theta|\mathbf{x})$, called the **log-likelihood function**, since it converts products into sums. Finding the value θ that maximizes the log-likelihood function ($\ln L(\theta|\mathbf{x})$) is equivalent to finding the value of θ that maximizes $L(\theta|\mathbf{x})$ since the natural logarithm is a monotonically increasing function. If $L(\theta|\mathbf{x})$ is differentiable with respect to θ , a possible mle is the solution to

$$\frac{\partial(\ln L(\theta|\mathbf{x}))}{\partial\theta} = 0. \quad (7.22)$$

Note that a possible mle is the solution to (7.22). A possible solution is used since a solution to (7.22) is a necessary but not sufficient condition for the solution to be a maximum, since the solution to (7.22) could be a local or global minimum, a local or global maximum, or a point of inflection. Recall that stationary points where,

$$\frac{\partial^2(\ln L(\theta|\mathbf{x}))}{\partial\theta^2} \Big|_{\theta=\hat{\theta}(\mathbf{x})} < 0, \quad (7.23)$$

indicate some type of maximum, either local or global. Further, the solution to (7.22) does not include points on the boundaries of the parameter space. Consequently, when evaluating the maximum of $L(\theta|\mathbf{x})$, the boundaries of the parameter space Θ as well as solutions to (7.22) must be evaluated.

Example 7.15 Given a random sample of size n taken from a *Bernoulli*(π) distribution, compute the maximum likelihood estimate and maximum likelihood estimator of the parameter π .

Solution: According to (4.2), the **pdf** for $X \sim \text{Bernoulli}(\pi)$ is

$$P(X = x|\pi) = \pi^x(1 - \pi)^{1-x},$$

where x takes on the value 1 with probability π and 0 with probability $1 - \pi$. The likelihood function for the n observed values is

$$L(\pi|\mathbf{x}) = \prod_{i=1}^n \pi^{x_i}(1 - \pi)^{1-x_i}.$$

Taking the natural logarithm of the likelihood function gives

$$\begin{aligned} \ln L(\pi|\mathbf{x}) &= \ln \left[\prod_{i=1}^n \pi^{x_i}(1 - \pi)^{1-x_i} \right] = \sum_{i=1}^n \ln [\pi^{x_i}(1 - \pi)^{1-x_i}] \\ &= \sum_{i=1}^n [x_i \ln \pi + (1 - x_i) \ln(1 - \pi)]. \end{aligned} \quad (7.24)$$

To find the value that maximizes (7.24), take the first-order partial derivative of $\ln L(\pi|\mathbf{x})$ with respect to π and set the answer equal to zero:

$$\frac{\partial \ln L(\pi|\mathbf{x})}{\partial \pi} = \frac{\sum_{i=1}^n x_i}{\pi} - \frac{n - \sum_{i=1}^n x_i}{1 - \pi} \stackrel{\text{set}}{=} 0. \quad (7.25)$$

The solution to (7.25) is $\pi = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}$. For $\pi = \bar{x}$ to be a maximum, the second-order partial derivative of the log-likelihood function must be negative at $\pi = \bar{x}$. The second-order partial derivative is

$$\frac{\partial^2 \ln L(\pi|\mathbf{x})}{\partial \pi^2} = \frac{-\sum_{i=1}^n x_i}{\pi^2} - \frac{n - \sum_{i=1}^n x_i}{(1 - \pi)^2}.$$

Evaluating the second-order partial derivative at $\pi = \bar{x}$ yields

$$\frac{\partial^2 \ln L(\pi|\mathbf{x})}{\partial \pi^2} = \frac{-n\bar{x}}{\bar{x}^2} - \frac{(n - n\bar{x})}{(1 - \bar{x})^2} = -\frac{n}{\bar{x}} - \frac{n}{1 - \bar{x}},$$

which is less than zero since $0 \leq \bar{x} \leq 1$ and $n > 0$. Finally, since the values of the likelihood function at the boundaries of the parameter space, $\pi = 0$ and $\pi = 1$, are 0, it follows that $\pi = \bar{x}$ is the value that maximizes the likelihood function. The maximum likelihood estimate $\hat{\pi}(\mathbf{x}) = \bar{x}$ and the maximum likelihood estimator $\hat{\pi}(\mathbf{X}) = \bar{X}$. ■

Example 7.16 ▷ *MLEs with S: Oriental Cockroaches* ◁ A laboratory is interested in testing a new child-friendly pesticide on *Blatta orientalis* (oriental cockroaches). The scientists from the lab apply the new pesticide to 81 randomly selected *Blatta orientalis oothecae* (eggs). The results from the experiment are stored in the data frame `Roacheggs` in the variable `eggs`. A zero in the variable `eggs` indicates that nothing hatched from the egg while a 1 indicates the birth of a cockroach. Assuming the selected *Blatta orientalis* eggs are representative of the population of *Blatta orientalis* eggs, estimate the proportion of *Blatta orientalis* eggs that result in a birth after being sprayed with the child-friendly pesticide. Use either `nlm()` in R or `nlmin()` in S-PLUS to solve the problem iteratively and to produce a graph of the log-likelihood function.

Solution: Note that whether or not a *Blatta orientalis* egg hatches is a Bernoulli trial with unknown parameter π . Using the maximum likelihood estimate from Example 7.15 on the facing page, $\hat{\pi}(\mathbf{x}) = \bar{x} = 0.21$.

```
> attach(Roacheggs)
> str(Roacheggs)          # Note: str(object) only works in R
'data.frame':   81 obs. of  1 variable:
 $ eggs: num  0 0 1 0 0 0 0 0 0 1 ...
> mean(eggs)
[1] 0.2098765
```

Both R and S-PLUS have iterative procedures that will minimize a given function. The minimization function in R is `nlm()`, while the minimization function in S-PLUS is `nlmin()`. The required arguments for both functions are `f()` and `p`, where `f()` is the function to be minimized and `p` is a vector of initial values for the parameter(s). Since both `nlm()` and `nlmin()` are minimization procedures and finding a maximum likelihood estimate is a maximization procedure, the functions `nlm()` and `nlmin()` on the negative of the log-likelihood function are used.

```
> p <- seq(0.1, 0.9, 0.001)
> negloglike <- function(p){-(sum(eggs)*log(p) + sum(1-eggs)*log(1-p))}
> nlm(negloglike, 0.2)
$minimum
[1] 41.61724

$estimate
[1] 0.2098760
```

```
$gradient
[1] 1.421085e-08
```

```
$code
[1] 1
```

```
$iterations
[1] 4
```

Warning messages:

```
1: In log(1 - p) : NaNs produced
2: In nlm(negloglike, 0.2) : NA/Inf replaced by maximum positive value
```

The following generic S code can be used to represent graphically the log-likelihood function in a fashion similar to Figure 7.3:

```
> par(pty = "s")
> p <- seq(0.1, 0.9, 0.001)
> plot(p, - negloglike(p), type = "n", ylab = "L")
> abline(v = mean(eggs), col = 13, lwd = 3)
> lines(p, - negloglike(p), col = 6, lwd = 3)
```

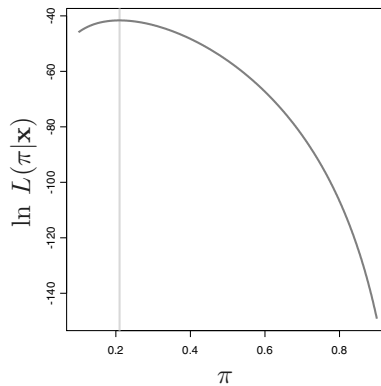


FIGURE 7.3: Illustration of the $\ln L(\pi|\mathbf{x})$ function for Example 7.16 ■

The function `optimize()`, available in both R and S-PLUS, approximates a local optimum of a continuous univariate function (`f`) within a given interval. The function searches the user-provided interval for either a minimum (default) or maximum of the function `f`. To solve Example 7.16 with `optimize()`, enter

```
> loglike <- function(p){(sum(eggs)*log(p) + sum(1-eggs) * log(1-p))}
> optimize(f=loglike, interval=c(0,1), maximum=TRUE)
$maximum
[1] 0.2098906
```

```
$objective
[1] -41.61724
```

Example 7.17 Let X_1, X_2, \dots, X_m be a random sample from a $Bin(n, \pi)$ population. Compute the maximum likelihood estimator and the maximum likelihood estimate for the parameter π . Verify your answer with simulation by generating 1000 random values from a $Bin(n = 3, \pi = 0.5)$ population.

Solution: The likelihood function is

$$\begin{aligned} L(\pi|\mathbf{x}) &= \prod_{i=1}^m \binom{n}{x_i} \pi^{x_i} (1 - \pi)^{n-x_i} \\ &= \binom{n}{x_1} \pi^{x_1} (1 - \pi)^{n-x_1} \times \dots \times \binom{n}{x_m} \pi^{x_m} (1 - \pi)^{n-x_m}, \end{aligned} \quad (7.26)$$

and the log-likelihood function is

$$\begin{aligned} \ln L(\pi|\mathbf{x}) &= \ln \left[\prod_{i=1}^m \binom{n}{x_i} \pi^{x_i} (1 - \pi)^{n-x_i} \right] = \sum_{i=1}^m \ln \left[\binom{n}{x_i} \pi^{x_i} (1 - \pi)^{n-x_i} \right] \\ &= \sum_{i=1}^m \left[\ln \binom{n}{x_i} + x_i \ln \pi + (n - x_i) \ln(1 - \pi) \right]. \end{aligned} \quad (7.27)$$

Next, look for the value that maximizes the log-likelihood function by taking the first-order partial derivative of (7.27) and setting the answer to zero:

$$\frac{\partial \ln L(\pi|\mathbf{x})}{\partial \pi} = \frac{\sum_{i=1}^m x_i}{\pi} - \frac{mn - \sum_{i=1}^m x_i}{1 - \pi} \stackrel{\text{set}}{=} 0. \quad (7.28)$$

The solution to (7.28) is $\pi = \frac{\sum_{i=1}^m x_i}{mn} = \frac{\bar{x}}{n}$. For $\pi = \frac{\bar{x}}{n}$ to be a maximum, the second-order partial derivative of the log-likelihood function must be negative at $\pi = \frac{\bar{x}}{n}$. The second-order partial derivative is

$$\frac{\partial^2 \ln L(\pi|\mathbf{x})}{\partial \pi^2} = \frac{-\sum_{i=1}^m x_i}{\pi^2} - \frac{mn - \sum_{i=1}^m x_i}{(1 - \pi)^2}.$$

Evaluating the second-order partial derivative at $\pi = \frac{\bar{x}}{n}$ and using the substitution $\sum_{i=1}^m x_i = m\bar{x}$ yields

$$\begin{aligned} \frac{\partial^2 \ln L(\pi|\mathbf{x})}{\partial \pi^2} &= -\frac{m\bar{x}}{\left(\frac{\bar{x}}{n}\right)^2} - \frac{mn - m\bar{x}}{\left(1 - \frac{\bar{x}}{n}\right)} \\ &= -\frac{mn^2}{\bar{x}} - \frac{m(n - \bar{x})}{\frac{(n - \bar{x})^2}{n^2}} = -\frac{mn^2}{\bar{x}} - \frac{mn^2}{n - \bar{x}} < 0. \end{aligned}$$

Finally, since the values of the likelihood function at the boundaries of the parameter space, $\pi = 0$ and $\pi = 1$, are 0, it follows that $\pi = \frac{\bar{x}}{n}$ is the value that maximizes the likelihood function. The maximum likelihood estimate $\hat{\pi}(\mathbf{x}) = \frac{\bar{x}}{n}$ and the maximum likelihood estimator $\hat{\pi}(\mathbf{X}) = \frac{\bar{X}}{n}$.

To simulate $\pi = \frac{\sum_{i=1}^m x_i}{mn} = \frac{\bar{x}}{n}$, generate 1000 random values from a $Bin(n = 3, \pi = 0.5)$ population. Pay particular attention to the fact that $n = 3$ and $m = 1000$.

Calculation of $\pi = \frac{\sum_{i=1}^m x_i}{mn}$

```
> set.seed(23)
> sum(rbinom(1000, 3, 0.5))/(1000 * 3)
[1] 0.5063333
```


Calculation of $\pi = \frac{\bar{x}}{n}$

```
> set.seed(23)
> mean(rbinom(1000, 3, 0.5))/3
[1] 0.5063333
```

■

Example 7.18 Let X_1, X_2, \dots, X_m be a random sample from a $Pois(\lambda)$ population. Compute the maximum likelihood estimator and the maximum likelihood estimate for the parameter λ . Verify your answer with simulation by generating 1000 random values from a $Pois(\lambda = 5)$ population.

Solution: The likelihood function is

$$L(\lambda|\mathbf{x}) = \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} = e^{-n\lambda} \prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!}, \quad (7.29)$$

and the log-likelihood function is

$$\ln L(\lambda|\mathbf{x}) = \ln \left[e^{-n\lambda} \prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!} \right] = -n\lambda + \sum_{i=1}^n x_i \ln \lambda - \sum_{i=1}^n \ln(x_i!). \quad (7.30)$$

Next, look for the value that maximizes the log-likelihood function by taking the first-order partial derivative of (7.30) and setting the answer to zero:

$$\frac{\partial \ln L(\lambda|\mathbf{x})}{\partial \lambda} = -n + \frac{\sum_{i=1}^n x_i}{\lambda} \stackrel{\text{set}}{=} 0. \quad (7.31)$$

The solution to (7.31) is $\lambda = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}$. For $\lambda = \bar{x}$ to be a maximum, the second-order partial derivative of the log-likelihood function must be negative at $\lambda = \bar{x}$. The second-order partial derivative is

$$\frac{\partial^2 \ln L(\lambda|\mathbf{x})}{\partial \lambda^2} = -\frac{\sum_{i=1}^n x_i}{\lambda^2}.$$

Evaluating the second-order partial derivative at $\lambda = \bar{x}$ yields

$$\frac{\partial^2 \ln L(\lambda|\mathbf{x})}{\partial \lambda^2} = -\frac{n\bar{x}}{\bar{x}^2} = -\frac{n}{\bar{x}} < 0.$$

Finally, since the values of the likelihood function at the boundaries of the parameter space, $\lambda = 0$ and $\lambda = \infty$, are 0, it follows that $\lambda = \bar{x}$ is the value that maximizes the likelihood function. The maximum likelihood estimate $\hat{\lambda}(\mathbf{x}) = \bar{x}$ and the maximum likelihood estimator $\hat{\lambda}(\mathbf{X}) = \bar{X}$.

To simulate $\hat{\lambda}(\mathbf{x}) = \bar{x}$, generate 1000 random values from a $Pois(\lambda = 5)$ population:

```
> set.seed(99)
> mean(rpois(1000, 5))
[1] 4.986
```

■

Example 7.19 A box contains five pieces of candy. Some of the candies are alcoholic, and some are not. In an attempt to estimate the proportion of alcoholic candies, a sample of size $n = 3$ is taken with replacement that results in (a, a, n) (two alcoholic candies and one non-alcoholic candy). Write out the maximum likelihood function and use it to select the maximum likelihood estimate of π , the true proportion of alcoholic candies.

Solution: The possible values for π are $\frac{0}{5}$, $\frac{1}{5}$, $\frac{2}{5}$, $\frac{3}{5}$, $\frac{4}{5}$, and $\frac{5}{5}$. Since there is at least one alcoholic candy and there is at least one non-alcoholic candy, the values $\pi = 0$ and $\pi = 1$ must be ruled out. In this case, the observed sample values are $\mathbf{x}=(a, a, n)$. The likelihood function is

$$\begin{aligned} L(\pi|\mathbf{x}) &= f(\mathbf{x}|\pi) \\ &= f(a|\pi) \times f(a|\pi) \times f(n|\pi). \end{aligned}$$

Box	π	$L(\pi a, a, n)$
aaaaan	$\frac{4}{5}$	$\frac{4}{5} \cdot \frac{4}{5} \cdot \frac{1}{5} = \frac{6}{125}$
aaann	$\frac{3}{5}$	$\frac{3}{5} \cdot \frac{3}{5} \cdot \frac{2}{5} = \frac{18}{125}$
aannn	$\frac{2}{5}$	$\frac{2}{5} \cdot \frac{2}{5} \cdot \frac{3}{5} = \frac{12}{125}$
annnn	$\frac{1}{5}$	$\frac{1}{5} \cdot \frac{1}{5} \cdot \frac{4}{5} = \frac{4}{125}$

Since the value $\pi = \frac{3}{5}$ maximizes the likelihood function, consider $\hat{\pi}(\mathbf{x}) = \frac{3}{5}$ to be the maximum likelihood estimate for the proportion of candies that are alcoholic. ■

Example 7.20 ▷ **General MLE** ◁ The random variable X can take on the values 0, 1, 2, and 3 with probabilities $\mathbb{P}(X = 0) = p^3$, $\mathbb{P}(X = 1) = (1 - p)p^2$, $\mathbb{P}(X = 2) = (1 - p)^2$, and $\mathbb{P}(X = 3) = 2p(1 - p)$, where $0 < p < 1$.

- Do the given probabilities for the random variable X satisfy the conditions for a probability distribution of X ?
- Find the maximum likelihood estimate for p if a random sample of size $n = 150$ resulted in a 0 twenty-four times, a 1 fifty-four times, a 2 thirty-two times, and a 3 forty times.
- Graph the log-likelihood function and determine its maximum using either the function `nlm()` or the function `nlmin()`.

Solution: The answers are as follows:

- For the distribution of X to be a valid **pdf**, it must satisfy the following two conditions:

- $p(x) \geq 0$ for all x .
- $\sum_x p(x) = 1$.

Condition (1) is satisfied since $0 < p < 1$. Condition (2) is also satisfied since

$$\begin{aligned} \sum_x p(x) &= p^3 + (1 - p)p^2 + (1 - p)^2 + 2p(1 - p) \\ &= p^3 + p^2 - p^3 + 1 + p^2 - 2p + 2p - 2p^2 = 1. \end{aligned}$$

- The likelihood function is

$$\begin{aligned} L(p|\mathbf{x}) &= [(p^3)]^{24} [(1 - p)p^2]^{54} [(1 - p)^2]^{32} [2p(1 - p)]^{40} \\ &= 2^{40} p^{220} (1 - p)^{158}, \end{aligned}$$

and the log-likelihood function is

$$\ln [L(p|\mathbf{x})] = 40 \ln 2 + 220 \ln p + 158 \ln(1 - p). \quad (7.32)$$

Next, look for the value that maximizes the log-likelihood function by taking the first-order partial derivative of (7.32) with respect to p and setting the answer equal to zero:

$$\frac{\partial \ln [L(p|\mathbf{x})]}{\partial p} = \frac{220}{p} - \frac{158}{1-p} \stackrel{\text{set}}{=} 0. \quad (7.33)$$

The solution to (7.33) is $p = 0.58$. In order for $p = 0.58$ to be a maximum, the second-order partial derivative of (7.32) with respect to p must be negative. Since

$$\frac{\partial^2 \ln [L(\pi|\mathbf{x})]}{\partial p^2} = -\frac{220}{p^2} - \frac{158}{(1-p)^2} < 0 \text{ for all } p,$$

this value is a global maximum. Therefore, the maximum likelihood estimate of p , $\hat{p}(\mathbf{x}) = 0.58$.

(c) Generic S code to graph the log-likelihood function depicted in Figure 7.4 is

```
> par(pty = "s")
> p <- seq(0.01, 0.99, 0.001)
> loglike <- function(p)
+   {40 * log(2) - 220 * log(p) - 158 * log(1 - p)}
> plot(p, - loglike(p), type = "n")
> lines(p, - loglike(p), col = 6, lwd = 3)
> abline(v = 0.58, col = 13, lwd = 3)
```

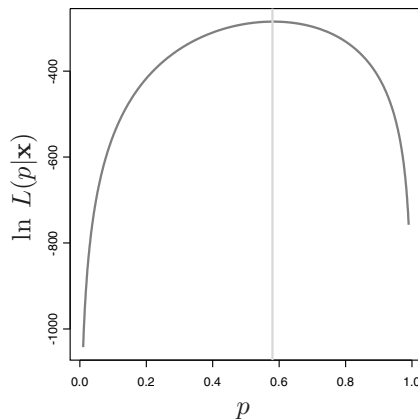


FIGURE 7.4: Illustration of the $\ln L(p|\mathbf{x})$ function for Example 7.20

To compute the maximum of the log-likelihood function, use the command `nlm(loglike, 0.001)` with R and the command `nlmin(loglike, 0.001)` with S-PLUS:

```
> nlm(loglike, 0.001)$estimate      # R
[1] 0.58201
```

Warning messages:

```
1: In log(1 - p) : NaNs produced
2: In nlm(loglike, 0.001) : NA/Inf replaced by maximum positive value
3: In log(1 - p) : NaNs produced
4: In nlm(loglike, 0.001) : NA/Inf replaced by maximum positive value
5: In log(1 - p) : NaNs produced
6: In nlm(loglike, 0.001) : NA/Inf replaced by maximum positive value
7: In log(1 - p) : NaNs produced
8: In nlm(loglike, 0.001) : NA/Inf replaced by maximum positive value
```

Example 7.21 A farmer cans and sells mild and hot peppers at the local market. The farmer recently hired an assistant to label his products. The assistant is new to working with peppers and has mislabeled some of the hot peppers as mild peppers. The farmer performs a random check of 100 of the mild pepper cans labeled by the assistant to assess his work. Out of the 100 cans labeled mild peppers, it turns out that 8 are actually hot peppers.

- (a) Which of the following proportions, 0.05, 0.08, or 0.10, maximizes the likelihood function?
- (b) What is the maximum likelihood estimate for the proportion of cans the assistant has mislabeled?

Solution: The answers are as follows:

(a) First define the random variable X as the number of mislabeled cans. In this definition of the random variable X , it follows that $n = 100$ and $m = 1$ since $X \sim \text{Bin}(100, \theta)$. The likelihood function for a random sample of size m from a $\text{Bin}(n, \pi)$ population was computed in (7.26) as

$$L(\pi|\mathbf{x}) = \prod_{i=1}^m \binom{n}{x_i} \pi^{x_i} (1 - \pi)^{n-x_i}.$$

Since $m = 1$ here, it follows that the likelihood function is

$$L(\pi|\mathbf{x}) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}.$$

Consequently, the value for π that maximizes

$$\mathbb{P}(X = 8|\pi) = \binom{100}{8} \pi^8 \cdot (1 - \pi)^{92}$$

is the solution to the problem. The likelihoods for the three values of π are

$$\mathbb{P}(X = 8|0.05) = \binom{100}{8} 0.05^8 \cdot (1 - 0.05)^{92} = 0.0648709,$$

$$\mathbb{P}(X = 8|0.08) = \binom{100}{8} 0.08^8 \cdot (1 - 0.08)^{92} = 0.1455185,$$

and

$$\mathbb{P}(X = 8|0.10) = \binom{100}{8} 0.10^8 \cdot (1 - 0.10)^{92} = 0.1148230.$$

Conclude that the value $\pi = 0.08$ is the value that maximizes the likelihood function among the three values of π provided.

(b) Recall that the maximum likelihood estimator for a binomial distribution was computed in Example 7.17 on page 261 as $\hat{\pi}(\mathbf{X}) = \frac{\sum_{i=1}^n x_i}{mn}$. Therefore, the maximum likelihood estimate for the proportion of mislabeled cans is $\hat{\pi}(\mathbf{x}) = \frac{8}{1 \cdot 100} = 0.08$. ■

Example 7.22 ▷ *I.I.D. Uniform Random Variables* ◁ Suppose $\{X_1, X_2, \dots, X_n\}$ is a random sample from a $Unif(0, \theta)$ distribution. Find the maximum likelihood estimator of θ . Find the maximum likelihood estimate for a randomly generated sample of 1000 $Unif(0, 3)$ random variables.

Solution: According to (4.9), the pdf of a random variable $X \sim Unif(0, \theta)$ is

$$f(x|\theta) = \frac{1}{\theta}, \quad 0 \leq x \leq \theta.$$

The likelihood function is

$$L(\theta|\mathbf{x}) = \begin{cases} \frac{1}{\theta^n} & \text{for } 0 \leq x_1 \leq \theta, 0 \leq x_2 \leq \theta, \dots, 0 \leq x_n \leq \theta \\ 0 & \text{otherwise.} \end{cases}$$

In this problem, the standard calculus approach fails since the maximum of the likelihood function occurs at a point of discontinuity. Consider the graph in Figure 7.5. Clearly $\frac{1}{\theta^n}$ is maximized for small values of θ . However, the likelihood function is only defined for $\theta \geq \max(x_i)$. Specifically, if $\theta < \max(x_i)$, $L(\theta|\mathbf{x}) = 0$. It follows then that the maximum likelihood estimator is $\hat{\theta}(\mathbf{X}) = \max(X_i)$. The following code finds the maximum likelihood estimate of 1000 randomly generated $Unif(0, 3)$ random variables:

```
> set.seed(2)
> max(runif(1000, 0, 3))
[1] 2.99781
```

Thus, even though a standard calculus approach could not be used, the mle 2.998667 is quite good for $\theta = 3$.

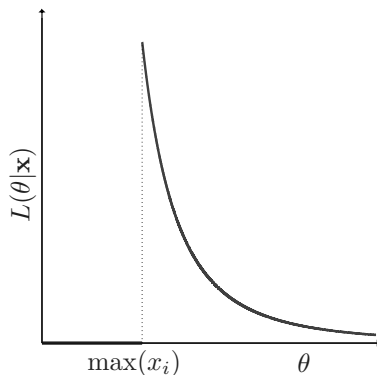


FIGURE 7.5: Illustration of the likelihood function in Example 7.22

Example 7.23 Suppose $\{X_1, X_2, \dots, X_n\}$ is a random sample from a $N(\mu, \sigma)$ distribution, where σ is assumed known. Find the maximum likelihood estimator of μ .

Solution: According to (4.23), the **pdf** of a random variable $X \sim N(\mu, \sigma)$ is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty.$$

The likelihood function is

$$L(\mu|\mathbf{x}) = \prod_{i=1}^n f(x_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}, \quad (7.34)$$

and the log-likelihood function is

$$\ln L(\mu|\mathbf{x}) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}. \quad (7.35)$$

To find the value of μ that maximizes $\ln L(\mu|\mathbf{x})$, take the first-order partial derivative of (7.35) with respect to μ , set the answer equal to zero, and solve. The first-order partial derivative of $\ln L(\mu|\mathbf{x})$ with respect to μ is

$$\frac{\partial \ln L(\mu, \sigma^2|\mathbf{x})}{\partial \mu} = \frac{\sum_{i=1}^n (x_i - \mu)}{\sigma^2} \stackrel{\text{set}}{=} 0. \quad (7.36)$$

The solution to (7.36) is $\mu = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}$. For $\mu = \bar{x}$ to be a maximum, the second-order partial derivative of the log-likelihood function with respect to μ must be negative at $\mu = \bar{x}$. The second-order partial derivative of (7.35) is

$$\frac{\partial^2 \ln L(\mu|\mathbf{x})}{\partial \mu^2} = -\frac{n}{\sigma^2} < 0. \quad (7.37)$$

Since (7.34) goes to zero at $\pm\infty$, the boundary values, it follows that $\mu = \bar{x}$ is a global maximum. Consequently, the maximum likelihood estimator of μ is $\hat{\mu}(\mathbf{X}) = \bar{X}$, and the maximum likelihood estimate of μ is $\hat{\mu}(\mathbf{x}) = \bar{x}$. ■

Example 7.24 Suppose $\{X_1, X_2, \dots, X_n\}$ is a random sample from a $N(\mu, \sigma)$ distribution, where μ is assumed known. Find the maximum likelihood estimator of σ^2 .

Solution: According to (4.23), the **pdf** of a random variable $X \sim N(\mu, \sigma)$ is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty.$$

The likelihood function is

$$L(\sigma^2|\mathbf{x}) = \prod_{i=1}^n f(x_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}, \quad (7.38)$$

and the log-likelihood function is

$$\ln L(\sigma^2|\mathbf{x}) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}. \quad (7.39)$$

To find the value of σ^2 that maximizes $\ln L(\sigma^2|\mathbf{x})$, take the first-order partial derivative of (7.39) with respect to σ^2 , set the answer equal to zero, and solve. The first-order partial derivative of $\ln L(\sigma^2|\mathbf{x})$ with respect to σ^2 is

$$\frac{\partial \ln L(\mu, \sigma^2|\mathbf{x})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^4} \stackrel{\text{set}}{=} 0. \quad (7.40)$$

The solution to (7.40) is $\sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n}$. For $\sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n}$ to be a maximum, the second-order partial derivative of the log-likelihood function with respect to σ^2 must be negative at $\sigma^2 = s_u^2$. For notational ease, let $r = \sigma^2$ in (7.39) so that

$$\ln L(r|\mathbf{x}) = \ln L(\sigma^2|\mathbf{x}) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(r) - \frac{\sum_{i=1}^n (x_i - \mu)^2}{2r}. \quad (7.41)$$

The second-order partial derivative of (7.41) is

$$\frac{\partial^2 \ln L(r|\mathbf{x})}{\partial r^2} = \frac{n}{2} r^{-2} - \sum_{i=1}^n (x_i - \mu)^2 r^{-3} \stackrel{?}{<} 0. \quad (7.42)$$

Multiplying the left-hand side of (7.42) by r^3 gives

$$\frac{n}{2} r - \sum_{i=1}^n (x_i - \mu)^2 \stackrel{?}{<} 0. \quad (7.43)$$

By substituting the value for the mle, $r = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n}$, the ? above the < can be removed since

$$\frac{r}{2} < \frac{\sum_{i=1}^n (x_i - \mu)^2}{n} = \sigma^2 = r.$$

Since (7.38) goes to zero at $\pm\infty$, the boundary values, it follows that $\sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n}$ is a global maximum. Consequently, the maximum likelihood estimator of σ^2 is $\widehat{\sigma}^2(\mathbf{X}) = \frac{\sum_{i=1}^n (X_i - \mu)^2}{n}$, and the maximum likelihood estimate of σ^2 is $\widehat{\sigma}^2(\mathbf{x}) = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n}$. ■

Example 7.25 Use `random.seed(33)` to generate 1000 $N(4, 1)$ random variables. Write log-likelihood functions for the simulated random variables and verify that the simulated maximum likelihood estimates for μ and σ^2 are reasonably close to the true parameters. Produce side-by-side graphs of $\ln L(\mu|\mathbf{x})$ and $\ln L(\sigma^2|\mathbf{x})$ indicating where the simulated maximum occurs in each graph.

Solution: The code provided is for R. For the given code to function in S-PLUS, replace the function `nlm()` with `nlmin()`.

```
> par(pty = "s")
> par(mfrow = c(1, 2))
> n <- 1000
> sigma <- 1
> set.seed(33)
> x <- rnorm(n, 4, sigma)
> mu <- seq(2, 6, length = n)
> negloglikemu <- function(mu)
+ { n/2 * log(2 * pi) + n/2 * log(sigma^2) + (sum(x^2)
+ - 2 * mu * sum(x) + n * mu^2)/(2 * sigma^2)}
```

```

> EM <- nlm(negloglikemu, 2)$estimate
> EM
[1] 4.019708
> mu1 <- 4
> negloglike <- function(sigma2)
+ {n/2 * log(2 * pi) + n/2 * log(sigma2) +
+ (sum((x - mu1)^2))/(2 * sigma2)}
> ES <- nlm(negloglike, 0.5)$estimate
Warning messages:
1: In log(sigma2) : NaNs produced
2: In nlm(negloglike, 0.5) : NA/Inf replaced by maximum positive value
> ES
[1] 1.000426

```

Note that the maximum likelihood estimates for μ and σ^2 from the simulation are 4.019708 and 1.000426, respectively, which are reasonably close to the parameters $\mu = 4$ and $\sigma^2 = 1$.

Code for the graph of $\ln L(\mu|\mathbf{x})$ versus μ is

```

> plot(mu, -negloglikemu(mu), type="n")
> lines(mu, -negloglikemu(mu), lwd=2)
> abline(v = EM, lty = 2)

```

Code for the graph of $\ln L(\sigma^2|\mathbf{x})$ versus σ^2 is

```

> sigma2 <- seq(0.5, 1.5, length = 1000)
> plot(sigma2, -negloglike(sigma2), type="n")
> lines(sigma2, -negloglike(sigma2), lwd=2)
> abline(v = ES, lty=2)

```

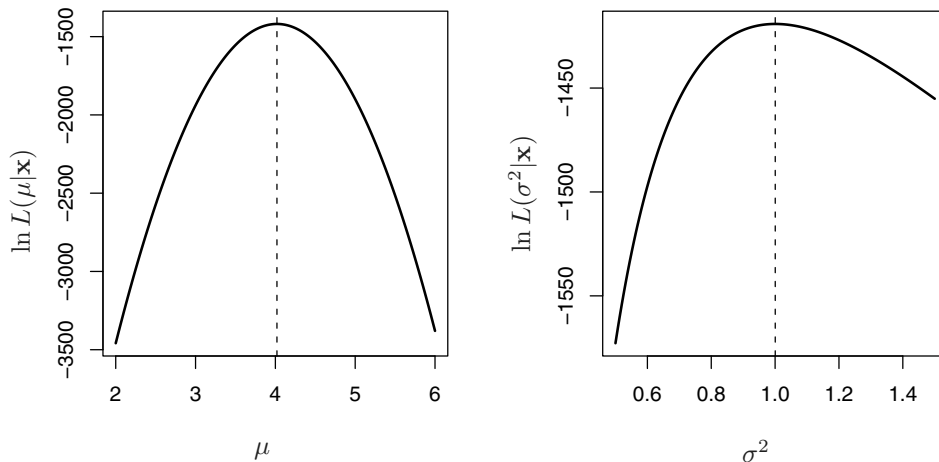


FIGURE 7.6: Illustration of $\ln L(\mu|\mathbf{x})$ and $\ln L(\sigma^2|\mathbf{x})$



7.3.2.1 Fisher Information

Now that proficiency has been gained at calculating point estimates and estimators with maximum likelihood procedures, some measure of the variance of these estimators is desired. Investigating a quantity known as the **Fisher information** or simply the **information number** will give this measure. The Fisher information is the amount of information that an observable random variable X carries about an unknown parameter θ , upon which the likelihood function of X , $L(\theta|\mathbf{x})$, depends. This is the quantity

$$E \left[\left(\frac{\partial \ln f(\mathbf{X}|\theta)}{\partial \theta} \right)^2 \right]. \quad (7.44)$$

This expression was briefly mentioned as the denominator of (7.6), the CRLB. However, the denominator of (7.6) used the form

$$n \cdot E \left[\left(\frac{\partial \ln f(X|\theta)}{\partial \theta} \right)^2 \right], \quad (7.45)$$

which is equivalent to (7.44) for random samples. Assume that X is a continuous random variable with **pdf** $f(x|\theta)$ (discrete random variables are handled in a similar fashion by exchanging integration for summation), where the following regularity conditions for $f(x|\theta)$ are satisfied:

1. The limits of support of $f(x|\theta)$ do not depend on θ .
2. The first two derivatives of $f(x|\theta)$ exist.
3. The order of integration and differentiation can be exchanged.

The inverse of the information number provides a bound for the variance of the best unbiased estimator of θ . Consequently, it makes sense to say the information number for a random sample of size n denoted $I_n(\theta)$ is the variance of the first-order partial derivative of the log-likelihood function. That is,

$$I_n(\theta) = \text{Var} \left[\left(\frac{\partial \ln f(\mathbf{X}|\theta)}{\partial \theta} \right) \right]. \quad (7.46)$$

When a random sample X_1, X_2, \dots, X_n is taken from a **pdf** $f(x|\theta)$, recall that $f(\mathbf{x}|\theta) = \prod_{i=1}^n f(x_i|\theta)$ so that $\ln f(\mathbf{x}|\theta) = \sum_{i=1}^n \ln f(x_i|\theta)$. When the random sample is of size $n = 1$, the Fisher information is denoted as simply $I(\theta)$, which is defined as

$$I(\theta) = \text{Var} \left[\left(\frac{\partial \ln f(X|\theta)}{\partial \theta} \right) \right]. \quad (7.47)$$

Since the random variables are independent, it should be clear that $I_n(\theta) = nI(\theta)$. The two common forms of expressing the information number for a random sample of size n are

$$I_n(\theta) = E \left[\left(\frac{\partial \ln f(\mathbf{X}|\theta)}{\partial \theta} \right)^2 \right] = nI(\theta) = nE \left[\left(\frac{\partial \ln f(X|\theta)}{\partial \theta} \right)^2 \right], \quad (7.48)$$

and

$$I_n(\theta) = -E \left[\left(\frac{\partial^2 \ln f(\mathbf{X}|\theta)}{\partial \theta^2} \right) \right] = nI(\theta) = -nE \left[\left(\frac{\partial^2 \ln f(X|\theta)}{\partial \theta^2} \right) \right]. \quad (7.49)$$

The form of the problem will often dictate which expression is easier to compute, as will be seen in the examples. The astute reader will have noticed that the equivalence of (7.48) and (7.49) was not shown, nor was the equivalence of (7.46) to (7.48) and (7.49).

Example 7.26 Given the **pdf** of a normal distribution with unknown mean μ and known variance σ^2 , find the Fisher information for μ using both (7.48) and (7.49) given a random sample of size n from said normal distribution.

Solution: According to (4.23), the **pdf** of a random variable $X \sim N(\mu, \sigma)$ is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty.$$

The likelihood function is

$$L(\mu|\mathbf{x}) = \prod_{i=1}^n f(x_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}},$$

and the log-likelihood function is

$$\ln L(\mu|\mathbf{x}) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}.$$

Note that

$$\frac{\partial \ln f(x|\mu)}{\partial \mu} = \frac{(x - \mu)}{\sigma^2},$$

and

$$\frac{\partial^2 \ln f(x|\mu)}{\partial \mu^2} = -\frac{1}{\sigma^2}.$$

Using (7.48), write

$$\begin{aligned} I_n(\mu) &= nE \left[\left(\frac{\partial \ln f(X|\mu)}{\partial \mu} \right)^2 \right] \\ &= nE \left[\left(\frac{X - \mu}{\sigma^2} \right)^2 \right] = n \frac{E[(X - \mu)^2]}{\sigma^4} = \frac{n\sigma^2}{\sigma^4} = \frac{n}{\sigma^2}. \end{aligned}$$

Using (7.49), write

$$\begin{aligned} I_n(\mu) &= -nE \left[\left(\frac{\partial^2 \ln f(X|\mu)}{\partial \mu^2} \right) \right] \\ &= -nE \left[-\frac{1}{\sigma^2} \right] = \frac{n}{\sigma^2}. \end{aligned}$$

Consequently, the smaller the variance σ^2 , the more information there is in a random sample of size n about μ . ■

7.3.2.2 Fisher Information for Several Parameters

Given a random variable X with **pdf** $f(x|\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$ is a k -dimensional vector of parameters, denote the information matrix of $\boldsymbol{\theta}$ as $\mathbf{I}(\boldsymbol{\theta})$. The (i, j) th element of the information matrix is defined as

$$I_{i,j}(\boldsymbol{\theta}) = E \left[\left(\frac{\partial \ln f(X|\boldsymbol{\theta})}{\partial \theta_i} \right) \left(\frac{\partial \ln f(X|\boldsymbol{\theta})}{\partial \theta_j} \right) \right] = E \left[\left(\frac{\partial^2 \ln f(X|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right) \right], \quad (7.50)$$

which is a generalization of (7.47). Likewise, when working with random samples,

$$\mathbf{I}_n(\boldsymbol{\theta}) = E \left[\left(\frac{\partial \ln f(\mathbf{X}|\boldsymbol{\theta})}{\partial \theta_i} \right) \left(\frac{\partial \ln f(\mathbf{X}|\boldsymbol{\theta})}{\partial \theta_j} \right) \right] = nI_{i,j}(\boldsymbol{\theta}), \quad (7.51)$$

and

$$\mathbf{I}_n(\boldsymbol{\theta}) = -E \left[\frac{\partial^2 \ln f(\mathbf{X}|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right] = nI_{i,j}(\boldsymbol{\theta}), \quad (7.52)$$

are the generalizations of (7.48) and (7.49), respectively.

Example 7.27 Given a random sample of size n from a $N(\mu, \sigma)$ population, where $\boldsymbol{\theta} = (\mu, \sigma^2)$, find $\mathbf{I}_n(\boldsymbol{\theta})$.

Solution: According to (4.23), the **pdf** of a random variable $X \sim N(\mu, \sigma)$ is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty.$$

It follows then that

$$f(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^n f(x_i|\boldsymbol{\theta}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}},$$

and that

$$\ln f(\mathbf{x}|\boldsymbol{\theta}) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}.$$

Taking partial derivatives of $\ln f(\mathbf{x}|\boldsymbol{\theta})$ with respect to $\theta_1 = \mu$, and $\theta_2 = \sigma^2$ gives

$$\begin{aligned} \frac{\partial \ln f(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_1} &= \frac{\sum_{i=1}^n (x_i - \mu)}{\sigma^2}, \\ \frac{\partial^2 \ln f(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_1} &= -\frac{n}{\sigma^2}, \\ \frac{\partial \ln f(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_2} &= -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n (x_i - \mu)^2}{2(\sigma^2)^2} \\ \frac{\partial^2 \ln f(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_2} &= \frac{n}{2(\sigma^2)^2} - \frac{\sum_{i=1}^n (x_i - \mu)^2}{(\sigma^2)^3}, \text{ and} \\ \frac{\partial^2 \ln f(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_2} &= \frac{\partial^2 \ln f(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_1} = -\frac{\sum_{i=1}^n (x_i - \mu)}{(\sigma^2)^2}. \end{aligned}$$

Using (7.52) gives

$$\mathbf{I}_n(\boldsymbol{\theta}) = -E \left[\frac{\partial^2 \ln f(\mathbf{X}|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right] = \begin{pmatrix} -E \left[\frac{\partial^2 \ln f(\mathbf{X}|\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_1} \right] & -E \left[\frac{\partial^2 \ln f(\mathbf{X}|\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_2} \right] \\ -E \left[\frac{\partial^2 \ln f(\mathbf{X}|\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_1} \right] & -E \left[\frac{\partial^2 \ln f(\mathbf{X}|\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_2} \right] \end{pmatrix},$$

or

$$\mathbf{I}_n(\boldsymbol{\theta}) = \begin{pmatrix} -E \left[-\frac{n}{\sigma^2} \right] & -E \left[-\frac{\sum_{i=1}^n (X_i - \mu)}{(\sigma^2)^2} \right] \\ -E \left[-\frac{\sum_{i=1}^n (X_i - \mu)}{(\sigma^2)^2} \right] & -E \left[\frac{n}{2(\sigma^2)^2} - \frac{\sum_{i=1}^n (X_i - \mu)^2}{(\sigma^2)^3} \right] \end{pmatrix},$$

which, upon taking expected values, becomes

$$\mathbf{I}_n(\boldsymbol{\theta}) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}. \quad \blacksquare$$

7.3.2.3 Properties of Maximum Likelihood Estimators

Now that the Fisher information has been examined and several problems have been worked with maximum likelihood estimation, the properties of maximum likelihood estimators are formally enumerated:

1. MLEs are not necessarily unbiased. For example, when sampling from a $N(\mu, \sigma)$ population, the MLE of σ^2 is $\widehat{\sigma}^2(\mathbf{X}) = \sum_{i=1}^n \frac{(X_i - \mu)^2}{n}$, which is a biased estimator of σ^2 . However, although some MLEs may be biased, all MLEs are consistent, which makes them asymptotically unbiased. Symbolically, MLEs $\not\Rightarrow$ unbiased estimators; however, MLEs \Rightarrow asymptotically unbiased estimators since MLEs \Rightarrow consistent estimators.
2. If T is a MLE of θ and g is any function, then $g(T)$ is the MLE of $g(\theta)$. This is known as the **invariance** property of MLEs. For example, if \bar{X} is the MLE of θ , then \bar{X}^2 is the MLE of θ^2 .
3. When certain regularity conditions on $f(x|\theta)$ are satisfied, and an efficient estimator exists for the estimated parameter, the efficient estimator is the MLE of the estimated parameter. Be careful, not all MLEs are efficient! However, if an efficient estimator exists, the efficient estimator is also the MLE. That is, efficiency \Rightarrow MLE, but MLE $\not\Rightarrow$ efficiency necessarily.
4. Under certain regularity conditions on $f(x|\theta)$, the MLE $\hat{\theta}(\mathbf{X})$ of θ based on a sample of size n from $f(x|\theta)$ is asymptotically normally distributed with mean θ and variance $I_n(\theta)^{-1}$. That is, as $n \rightarrow \infty$,

$$\hat{\theta}(\mathbf{X}) \sim N\left(\theta, \sqrt{I_n(\theta)^{-1}}\right). \quad (7.53)$$

The statement in (7.53) is the basis for large sample hypothesis tests (covered in Chapter 9) and confidence intervals (covered in Chapter 8).

Note that the asymptotic variance of MLEs equals the Cramér-Rao lower bound since they are asymptotically efficient. That is, MLEs \Rightarrow asymptotic efficiency. Consequently, a reasonable approximation to the distribution of $\hat{\theta}(\mathbf{X})$ for large sample sizes can be obtained. However, a normal distribution for $\hat{\theta}(\mathbf{X})$ cannot be guaranteed when the sample size is small.

Example 7.28 In Example 7.17 on page 261, it was found that the sample proportion of successes for a random sample of size m from a $Bin(n, \pi)$ distribution had $\hat{\pi} = \frac{\sum_{i=1}^m x_i}{mn}$ for its mle. That is, the MLE for the binomial proportion π is $\hat{\pi}(\mathbf{X}) = \frac{\sum_{i=1}^m X_i}{mn}$. What is the MLE for the variance of the sample proportion of successes where the random variable $\hat{\pi}$ is defined as $\frac{\sum_{i=1}^m X_i}{mn}$?

Solution: Given that $X \sim Bin(n, \pi)$, the variance of X is $n\pi(1 - \pi)$. Therefore,

$$\text{Var}[\hat{\pi}] = \text{Var}\left[\frac{\sum_{i=1}^m X_i}{mn}\right] = \frac{\sum_{i=1}^m \text{Var}[X_i]}{m^2 n^2} = \frac{mn\pi(1 - \pi)}{m^2 n^2} = \frac{\pi(1 - \pi)}{mn}.$$

Since $\text{Var}[\hat{\pi}]$ is a function of the MLE $\hat{\pi}(\mathbf{X})$, it follows using the invariance property of MLEs that the MLE of the variance of $\hat{\pi}$ is

$$\widehat{\text{Var}}[\hat{\pi}(\mathbf{X})] = \frac{\hat{\pi}(1 - \hat{\pi})}{mn}.$$

Note: Many texts will list the MLE of the variance of the sample proportion of successes in a binomial distribution as $\frac{\hat{\pi}(1 - \hat{\pi})}{n}$ because they use $m = 1$ in their definition of $\hat{\pi}$. ■

Example 7.29 ▷ *MOM and MLE for a Gamma* ◁ Given a random sample of size n from a population with pdf

$$f(x|\theta) = \frac{x}{\theta^2} e^{-\frac{x}{\theta}}, \quad x \geq 0, \quad \theta > 0,$$

- Find an estimator of θ using the method of moments.
- Find an estimator of θ using the method of maximum likelihood.
- Are the method of moments and maximum likelihood estimators of θ unbiased?
- Compute the variance of the MLE of θ .
- Is the MLE of θ efficient?

Solution: Since $X \sim \text{Gamma}(\alpha = 2, \lambda = \frac{1}{\theta})$, according to (4.16), $E[X] = \frac{\alpha}{\lambda} = 2\theta$ and $\text{Var}[X] = \frac{\alpha}{\lambda^2} = 2\theta^2$.

(a) Equating the first population moment about the origin to the first sample moment about the origin gives

$$\alpha_1(\theta) = 2\theta \stackrel{\text{set}}{=} \bar{X} = m_1,$$

which implies that the method of moments estimator for θ is $\tilde{\theta} = \frac{\bar{X}}{2}$.

(b) The likelihood equation is given as

$$L(\theta|\mathbf{x}) = \prod_{i=1}^n f(x_i) = \frac{\prod_{i=1}^n x_i}{\theta^{2n}} e^{-\frac{\sum_{i=1}^n x_i}{\theta}}, \quad (7.54)$$

and the log-likelihood function is

$$\ln L(\theta|\mathbf{x}) = -2n \ln(\theta) + \sum_{i=1}^n \ln(x_i) - \frac{\sum_{i=1}^n x_i}{\theta}. \quad (7.55)$$

To find the value of θ that maximizes $\ln L(\theta|\mathbf{x})$, take the first-order partial derivative of (7.55) with respect to θ , set the answer equal to zero, and solve. The first-order partial derivative of $\ln L(\theta|\mathbf{x})$ with respect to θ is

$$\frac{\partial \ln L(\theta|\mathbf{x})}{\partial \theta} = -\frac{2n}{\theta} + \frac{\sum_{i=1}^n x_i}{\theta^2} \stackrel{\text{set}}{=} 0. \quad (7.56)$$

The solution to (7.56) is $\theta = \frac{\bar{X}}{2}$, which agrees with the method of moments estimator. However, to ensure that $\theta = \frac{\bar{X}}{2}$ is a maximum, the second-order partial derivative with respect to θ must be negative. The second-order partial derivative of (7.55) is

$$\frac{\partial^2 \ln L(\theta|\mathbf{x})}{\partial \theta^2} = \frac{2n}{\theta^2} - \frac{2 \sum_{i=1}^n x_i}{\theta^3} \stackrel{?}{<} 0. \quad (7.57)$$

By using $\theta = \frac{\bar{X}}{2}$ in (7.57), arrive at the expression

$$-\frac{12n}{\bar{X}^2} \stackrel{?}{<} 0. \quad (7.58)$$

The ? above the < in (7.58) can be removed since $\int_0^\infty f(x) dx = 1 \Rightarrow \bar{X} > 0$. Finally, since (7.54) goes to zero as $\theta \rightarrow \infty$, it can be concluded that $\theta = \frac{\bar{X}}{2}$ is a global maximum. Consequently, the maximum likelihood estimator of θ is $\hat{\theta}(\mathbf{X}) = \frac{\bar{X}}{2}$.

(c) Since both the method of moments and the method of maximum likelihood returned the same estimator for θ , that is, $\hat{\theta}(\mathbf{X}) = \tilde{\theta} = \frac{\bar{X}}{2}$, the question is

$$E[\hat{\theta}(\mathbf{X})] = E[\tilde{\theta}] \stackrel{?}{=} \theta.$$

Both $\tilde{\theta}$ and $\hat{\theta}(\mathbf{X})$ are therefore unbiased estimators since

$$E[\hat{\theta}(\mathbf{X})] = E[\tilde{\theta}] = E\left[\frac{\bar{X}}{2}\right] = \frac{\sum_{i=1}^n E[X_i]}{2n} = \frac{n \cdot 2\theta}{2n} = \theta.$$

(d) The variance of the MLE of θ is

$$\text{Var}[\hat{\theta}(\mathbf{X})] = \text{Var}\left[\frac{\bar{X}}{2}\right] = \text{Var}\left[\frac{\sum_{i=1}^n X_i}{2n}\right] = \frac{n \text{Var}[X]}{4n^2} = \frac{n2\theta^2}{4n^2} = \frac{\theta^2}{2n}.$$

(e) For $\hat{\theta}(\mathbf{X}) = \frac{\bar{X}}{2}$ to be considered an efficient or minimum variance estimator of θ , the variance of $\frac{\bar{X}}{2}$ must equal the CRLB. That is, does

$$\text{Var}[\hat{\theta}(\mathbf{X})] = \frac{\theta^2}{2n} \stackrel{?}{=} \frac{1}{n \cdot E\left[\left(\frac{\partial \ln f(X|\theta)}{\partial \theta}\right)^2\right]}?$$

Since $f(x|\theta) = \frac{x}{\theta^2} e^{-\frac{x}{\theta}}$ for $x \geq 0$, and $\theta > 0$, it follows that $\ln f(x|\theta) = \ln x - 2 \ln \theta - \frac{x}{\theta}$, and that $\frac{\partial \ln f(x|\theta)}{\partial \theta} = \frac{x-2\theta}{\theta^2}$. Consequently,

$$\frac{1}{n \cdot E\left[\left(\frac{\partial \ln f(X|\theta)}{\partial \theta}\right)^2\right]} = \frac{1}{n \cdot E\left[\left(\frac{X-2\theta}{\theta^2}\right)^2\right]} = \frac{1}{\frac{n \cdot \text{Var}[X]}{\theta^4}} = \frac{1}{\frac{n \cdot 2\theta^2}{\theta^4}} = \frac{\theta^2}{2n},$$

and conclude that $\frac{\bar{X}}{2}$ is an efficient estimator of θ . ■

Example 7.30 ▷ *MLEs for Exponentials* ◁ Given a random sample of size n from an exponential distribution with pdf

$$f(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}} \quad x \geq 0, \quad \theta > 0, \quad (7.59)$$

- Find the MLE of θ^2 .
- Show that the MLE of θ^2 is a biased estimator of θ^2 .
- Provide an unbiased estimator of θ^2 .
- Find the variance of your MLE of θ^2 .
- Find the variance of your unbiased estimator of θ^2 .
- Show that the variance for the MLE of θ^2 converges to $I_n(\theta)^{-1}$ as $n \rightarrow \infty$ according to property 4 of the Properties of MLEs on page 273.

Solution: To find the MLE of θ^2 , there are two possibilities. First, the MLE of θ could be found and the invariance property could be used to say that this estimate squared is the MLE of θ^2 . (See problem 37 of this chapter.) Second, and this is the current approach, the MLE of θ^2 can be found directly.

(a) For notational ease, use the change of variable $\theta^2 = p$ and $\theta = \sqrt{p}$ in (7.59). The resulting **pdf** using the change of variable is

$$f(x) = \frac{1}{\sqrt{p}} e^{-\frac{x}{\sqrt{p}}} \quad x \geq 0, \quad p > 0.$$

The likelihood function is

$$L(p|\mathbf{x}) = \prod_{i=1}^n f(x_i) = \prod_{i=1}^n \frac{1}{\sqrt{p}} e^{-\frac{x_i}{\sqrt{p}}} = \frac{1}{(\sqrt{p})^n} e^{-\frac{\sum_{i=1}^n x_i}{\sqrt{p}}}, \quad (7.60)$$

and the log-likelihood function is

$$\ln L(p|\mathbf{x}) = -\frac{n}{2} \ln p - \frac{\sum_{i=1}^n x_i}{\sqrt{p}}. \quad (7.61)$$

To find the value of p that maximizes $\ln L(p|\mathbf{x})$, take the first-order partial derivative of (7.61) with respect to p , set the answer equal to zero, and solve. The first-order partial derivative of $\ln L(p|\mathbf{x})$ with respect to p is

$$\frac{\partial \ln L(p|\mathbf{x})}{\partial p} = -\frac{n}{2p} + \frac{\sum_{i=1}^n x_i}{2p^{\frac{3}{2}}} \stackrel{\text{set}}{=} 0. \quad (7.62)$$

The solution to (7.62) is $p = \bar{x}^2$. For $p = \bar{x}^2$ to be a maximum, the second-order partial derivative of the log-likelihood function with respect to p must be negative at $p = \bar{x}^2$. The second-order partial derivative of (7.61) is

$$\frac{\partial^2 \ln L(p|\mathbf{x})}{\partial p^2} = \frac{n}{2p^2} - \frac{3 \sum_{i=1}^n x_i}{4p^{\frac{5}{2}}} \stackrel{?}{<} 0. \quad (7.63)$$

By substituting $p = \bar{x}^2$ in the right-hand side of (7.63), the ? above the < can be removed since $\bar{x} < \frac{3\bar{x}}{2}$ because $\bar{x} > 0$ for any sample due to the fact that $\mathbb{P}(X = 0) = 0$ for any continuous distribution. Finally, since as $p \rightarrow \infty$, $L(p|\mathbf{x}) \rightarrow 0$, it can be concluded that the MLE of $p = \theta^2$ is $\hat{p}(\mathbf{X}) = \hat{\theta}^2(\mathbf{X}) = \bar{X}^2$.

(b) Next, show that \bar{X}^2 is a biased estimator of θ^2 . The easiest way to determine the mean and variance of \bar{X}^2 is with moment generating functions. It is known that the moment generating function of an exponential random variable, X , is $M_X(t) = (1 - \theta t)^{-1}$. Furthermore, if $Y = \sum_{i=1}^n c_i X_i$ and each X_i has a moment generating function $M_{X_i}(t)$, then the moment generating function of Y is $M_Y(t) = \prod_{i=1}^n M_{X_i}(c_i t)$. In the case where $Y = \bar{X} = \frac{\sum_{i=1}^n X_i}{n}$, each $c_i = \frac{1}{n}$. For the special case of the exponential, the moment generating function for \bar{X} is

$$M_{\bar{X}}(t) = M_Y(t) = \prod_{i=1}^n \left(1 - \theta \cdot \frac{t}{n}\right)^{-1} = \left(1 - \frac{\theta t}{n}\right)^{-n}.$$

Thus, to calculate the mean and variance of \bar{X}^2 , take the first through fourth derivatives of $M_{\bar{X}}(t)$ and evaluate them when $t = 0$ to find $E[\bar{X}^i]$ for $i = 1, 2, 3$, and 4. The first,

second, third, and fourth derivatives of $M_{\bar{X}}(t)$, respectively, are

$$\begin{aligned} M'_{\bar{X}}(t) &= -n \left(1 - \frac{\theta}{n}t\right)^{-n-1} \left(-\frac{\theta}{n}\right) \\ M''_{\bar{X}}(t) &= \theta(-n-1) \left(1 - \frac{\theta}{n}t\right)^{-n-2} \left(-\frac{\theta}{n}\right) \\ M'''_{\bar{X}}(t) &= \frac{\theta^2(n+1)}{n}(-n-2) \left(1 - \frac{\theta}{n}t\right)^{-n-3} \left(-\frac{\theta}{n}\right) \\ M^{(4)}_{\bar{X}}(t) &= \frac{\theta^3(n+1)(n+2)}{n^2}(-n-3) \left(1 - \frac{\theta}{n}t\right)^{-n-4} \left(-\frac{\theta}{n}\right) \end{aligned}$$

Evaluating these derivatives at $t = 0$ gives the expected values of \bar{X} to the first, second, third, and fourth powers:

$$\begin{aligned} M'_{\bar{X}}(0) &= \theta = E[\bar{X}] \\ M''_{\bar{X}}(0) &= \frac{\theta^2(n+1)}{n} = E[\bar{X}^2] \\ M'''_{\bar{X}}(0) &= \frac{\theta^3(n+1)(n+2)}{n^2} = E[\bar{X}^3] \\ M^{(4)}_{\bar{X}}(0) &= \frac{\theta^4(n+1)(n+2)(n+3)}{n^3} = E[\bar{X}^4] \end{aligned}$$

Since $E[\bar{X}^2] = \frac{\theta^2(n+1)}{n} \neq \theta^2$, \bar{X}^2 is a biased estimator of θ^2 .

(c) An unbiased estimator of θ^2 would be to use the quantity $\frac{n\bar{X}^2}{n+1}$.

(d) The variance of \bar{X}^2 can be computed as $E[\bar{X}^4] - (E[\bar{X}^2])^2$:

$$\begin{aligned} \text{Var}[\bar{X}^2] &= \frac{\theta^4(n+1)(n+2)(n+3)}{n^3} - \left(\frac{\theta^2(n+1)}{n}\right)^2 \\ &= \frac{2\theta^4(2n^2 + 5n + 3)}{n^3} \\ &= \frac{2\theta^4((2n+3)(n+1))}{n^3} \end{aligned} \tag{7.64}$$

(e) The variance of the unbiased estimator of θ^2 is

$$\begin{aligned} \text{Var}\left[\frac{n\bar{X}^2}{n+1}\right] &= \frac{n^2}{(n+1)^2} \text{Var}[\bar{X}^2] \\ &= \frac{n^2}{(n+1)^2} \cdot \frac{2\theta^4((2n+3)(n+1))}{n^3} \\ &= \frac{2\theta^4(2n+3)}{n(n+1)}. \end{aligned}$$

(f) The Fisher information is computed as

$$\begin{aligned} I_n(p) &= -E \left[\left(\frac{\partial^2 \ln f(\mathbf{X}|p)}{\partial p^2} \right) \right] = -E \left[\frac{n}{2p^2} - \frac{3 \sum_{i=1}^n x_i}{4p^{\frac{5}{2}}} \right] \\ &= - \left[\frac{n}{2p^2} - \frac{3n\sqrt{p}}{4p^{\frac{5}{2}}} \right] = \frac{n}{2p^2} \left[-1 + \frac{3}{2} \right] = \frac{n}{4p^2}. \end{aligned}$$

Since $p = \theta^2$, it follows that $I_n(p) = I_n(\theta^2) = \frac{n}{4\theta^4}$, and that $I_n(\theta^2)^{-1} = \frac{4\theta^4}{n}$. Note that the variance of the MLE estimator \bar{X}^2 given in (7.64) converges to $I_n(\theta^2)^{-1} = \frac{4\theta^4}{n}$ as $n \rightarrow \infty$. ■

7.3.2.4 Finding Maximum Likelihood Estimators for Multiple Parameters

When the pdf contains more than one parameter, the procedure for finding the MLEs for several parameters proceeds in a fashion analogous to the one-parameter case. Given a vector $\boldsymbol{\theta} = \{\theta_1, \theta_2, \dots, \theta_k\}$, the likelihood function is represented as

$$\begin{aligned} L(\boldsymbol{\theta}|\mathbf{x}) &= L(\theta_1, \dots, \theta_k | x_1, \dots, x_n) \\ &= f(x_1 | \theta_1, \dots, \theta_k) \times \dots \times f(x_n | \theta_1, \dots, \theta_k). \end{aligned} \quad (7.65)$$

The value of $\boldsymbol{\theta}$ that maximizes $L(\boldsymbol{\theta}|\mathbf{x})$ is the mle of $\boldsymbol{\theta}$. In the multiple parameter case, denote the mle of $\boldsymbol{\theta}$ as $\hat{\boldsymbol{\theta}}(\mathbf{x})$ and the MLE of $\boldsymbol{\theta}$ as $\hat{\boldsymbol{\theta}}(\mathbf{X})$. As with the univariate case, typically work with the log-likelihood function ($\ln L(\boldsymbol{\theta}|\mathbf{x})$) instead of the likelihood function. If $L(\boldsymbol{\theta}|\mathbf{x})$ is differentiable with respect to $\boldsymbol{\theta}$, a possible mle for $\boldsymbol{\theta}$ are the θ_i s, $i = 1, \dots, k$, that solve

$$\frac{\partial \ln L(\boldsymbol{\theta}|\mathbf{x})}{\partial \boldsymbol{\theta}} \stackrel{\text{set}}{=} \mathbf{0} \Leftrightarrow \begin{cases} \frac{\partial \ln L(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_1} = \sum_{i=1}^n \frac{\partial \ln f(x_i|\boldsymbol{\theta})}{\partial \theta_1} \stackrel{\text{set}}{=} 0 \\ \vdots \\ \frac{\partial \ln L(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_k} = \sum_{i=1}^n \frac{\partial \ln f(x_i|\boldsymbol{\theta})}{\partial \theta_k} \stackrel{\text{set}}{=} 0. \end{cases} \quad (7.66)$$

Just as with the univariate case, possible mles for $\boldsymbol{\theta}$ are the solutions to (7.66). Solutions to the k equations in (7.66) are a necessary but not sufficient condition for the solutions to be maximums. However, a sufficient condition to guarantee the solutions to (7.66) are maxima is for the Hessian matrix (matrix whose elements are the second-order partial derivatives with respect to the parameters being estimated) to be negative definite when evaluated at the maximum likelihood estimators. Any symmetric $p \times p$ matrix is **negative definite** provided the leading principal minors (the determinants of the upper left square submatrices) have alternating signs where the top left element in the matrix is negative. These principal minors are denoted by D_i for $i = 1, \dots, p$ and satisfy the following conditions: $D_1 < 0$, $D_2 > 0, \dots$, ending with $D_p > 0$ if p is even and $D_p < 0$ if p is odd (Casella and Berger, 1990). Furthermore, the solutions to (7.66) will yield minima when the determinants of the leading principal minors are all positive.

Example 7.31 Given a random sample of size n from a normal distribution with unknown mean μ and variance σ^2 , find the MLEs for μ and σ^2 .

Solution: The pdf for a random variable $X \sim N(\mu, \sigma)$ according to (4.23) is

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

The likelihood function is

$$L(\mu, \sigma^2|\mathbf{x}) = \prod_{i=1}^n f(x_i) = \frac{1}{(2\pi)^{\frac{n}{2}} (\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2},$$

and the log-likelihood function is

$$\ln L(\mu, \sigma^2|\mathbf{x}) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2. \quad (7.67)$$

To find the $\boldsymbol{\theta}$ that maximizes (7.67), take the first-order partial derivatives with respect to $\boldsymbol{\theta} = (\mu, \sigma^2)$, set those first-order partial derivatives equal to zero, and solve the simultaneous equations:

$$\frac{\partial \ln L(\boldsymbol{\theta}|\mathbf{x})}{\partial \boldsymbol{\theta}} \stackrel{\text{set}}{=} \mathbf{0} \Leftrightarrow \begin{cases} \frac{\partial \ln L(\mu, \sigma^2|\mathbf{x})}{\partial \mu} = \frac{\sum_{i=1}^n (x_i - \mu)}{\sigma^2} \stackrel{\text{set}}{=} 0 \\ \frac{\partial \ln L(\mu, \sigma^2|\mathbf{x})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^4} \stackrel{\text{set}}{=} 0. \end{cases}$$

The solution to the system of equations is

$$\mu = \bar{x} \quad \text{and} \quad \sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}.$$

A sufficient condition for the values in $\boldsymbol{\theta}$ to be maximums is for the Hessian matrix to be negative definite. In this case, the Hessian matrix is

$$H = \begin{pmatrix} \frac{\partial^2 \ln L(\mu, \sigma^2|\mathbf{x})}{\partial \mu^2} & \frac{\partial^2 \ln L(\mu, \sigma^2|\mathbf{x})}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 \ln L(\mu, \sigma^2|\mathbf{x})}{\partial \sigma^2 \partial \mu} & \frac{\partial^2 \ln L(\mu, \sigma^2|\mathbf{x})}{\partial (\sigma^2)^2} \end{pmatrix}.$$

Specifically, the second-order partial derivatives are

$$\begin{aligned} \frac{\partial^2 \ln L(\mu, \sigma^2|\mathbf{x})}{\partial \mu^2} &= -\frac{n}{\sigma^2}, \\ \frac{\partial^2 \ln L(\mu, \sigma^2|\mathbf{x})}{\partial (\sigma^2)^2} &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^n (x_i - \mu)^2, \quad \text{and} \\ \frac{\partial^2 \ln L(\mu, \sigma^2|\mathbf{x})}{\partial \mu \partial \sigma^2} &= -\frac{1}{\sigma^4} \sum_{i=1}^n (x_i - \mu). \end{aligned}$$

By substituting the values $\mu = \bar{x}$ and $\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = s_u^2$ in the second-order partial derivatives, the Hessian matrix is expressed as

$$H = \begin{pmatrix} -\frac{n}{s_u^2} & 0 \\ 0 & -\frac{n}{2s_u^4} \end{pmatrix}.$$

Note that H is negative definite since $D_1 = -\frac{n}{s_u^2} < 0$ and $D_2 = \frac{n^2}{2s_u^6} > 0$, implying that the solutions, $\mu = \bar{x}$ and $\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$, are maximums. Finally, the solutions $\mu = \bar{x}$ and $\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$ can be considered global maximums since the likelihood function goes to zero for both $\mu = \pm\infty$ and $\sigma^2 = \infty$. Consequently, the MLE of θ is written as $\hat{\theta}(\mathbf{X}) = (\bar{X}, S_u^2)$, and the mle of θ as $\hat{\theta}(\mathbf{x}) = (\bar{x}, s_u^2)$.

Example 7.32 Use `set.seed(11)` to generate 500 values from a $N(2, 1)$ population, and treat the generated values as a random sample of size $n = 500$ from a normal distribution with unknown parameters. Find the maximum likelihood estimates for μ and σ^2 based on the generated sample.

Solution: According to the results of Example 7.31 on page 278, the MLE of θ when sampling from a normal distribution with unknown mean and variance is $\hat{\theta}(\mathbf{X}) = (\bar{X}, S_u^2)$. The following S code performs the simulation:

```
> set.seed(11)
> n <- 500
> x <- rnorm(n, 2, 1)
> mean(x)
[1] 1.997360
> S2u <- sum((x - mean(x))^2/n)
> S2u
[1] 0.9764792
```

From this simulation, $\hat{\theta}(\mathbf{x}) = (1.997360, 0.9764792)$. Another approach is to allow S to find the values that maximize the log-likelihood function analytically using either `nlm()` or `nlmin()`. The code that follows is for R. To compute the answer with S-PLUS, replace the command `nlm(negloglike, c(3,2))$estimate` with `nlmin(negloglike, c(3,2))`.

```
> negloglike <- function(p)
+ { (n/2)*log(2*pi) + (n/2)*log(2*p[2]) + (1/(2*p[2]))*sum((x - p[1])^2) }
> nlm(negloglike, c(3, 2))$estimate
[1] 1.9973587 0.9764787
```

Warning messages:

```
1: In log(2 * p[2]) : NaNs produced
2: In nlm(negloglike, c(3, 2)) : NA/Inf replaced by maximum positive value
```

7.3.2.5 Multi-Parameter Properties of MLEs

The four properties for a MLE $\hat{\theta}(\mathbf{X})$ of θ given in Section 7.3.2.3 on page 273 also apply to a k -dimensional vector $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ of parameters. Of particular importance is the generalization of property 4 on page 273. Specifically, property 4 on page 273 states that, under certain regularity conditions on $f(x|\theta)$, the MLE of $\hat{\theta}(\mathbf{X})$ of θ based on a sample of size n is asymptotically normally distributed with mean θ and variance-covariance matrix $I_n(\theta)^{-1}$. That is,

$$\hat{\theta} \sim N(\theta, I_n(\theta)^{-1}),$$

and the variance-covariance MLEs are

$$\widehat{I_n(\theta)}^{-1} = I_n(\theta)^{-1}|_{\theta=\hat{\theta}(\mathbf{X})}.$$

Example 7.33 Given a random sample of size n from a $N(\mu, \sigma)$ population, find the MLE of the variance of \bar{X} and the variance of S_u^2 .

Solution: In Example 7.31 on page 278, the MLE of $\boldsymbol{\theta}$ was $\hat{\boldsymbol{\theta}}(\mathbf{X}) = (\bar{X}, S_u^2)$, and in Example 7.27 on page 272, the Fisher information matrix was

$$\mathbf{I}_n(\boldsymbol{\theta}) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}.$$

Consequently,

$$\begin{aligned} \widehat{\mathbf{I}}_n(\widehat{\boldsymbol{\theta}})^{-1} &= \mathbf{I}_n(\boldsymbol{\theta})^{-1} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{X})} = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}^{-1} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{X})} \\ &= \begin{pmatrix} \frac{n}{S_u^2} & 0 \\ 0 & \frac{n}{2S_u^4} \end{pmatrix}^{-1} = \begin{pmatrix} \frac{S_u^2}{n} & 0 \\ 0 & \frac{2S_u^4}{n} \end{pmatrix}, \end{aligned}$$

from which it can be concluded that

$$I_{11}(\widehat{\boldsymbol{\theta}})^{-1} = \widehat{\text{Var}}(\bar{X}) = \frac{S_u^2}{n},$$

and

$$I_{22}(\widehat{\boldsymbol{\theta}})^{-1} = \widehat{\text{Var}}(S_u^2) = \frac{2S_u^4}{n}. \quad \blacksquare$$

7.4 Problems

- Use the data from the data frame `WheatSpain` to answer the questions.
 - Find the mean, median, standard deviation, and *MAD* of the `wheat.surface`.
 - Remove the `Castilla-Leon` community and find again the mean, median, standard deviation, and *MAD* of the same variable. Which statistics are preferred as measures for these data? Comment on the results.
- Given the estimators of the mean $T_1 = (X_1 + 2X_2 + X_3)/4$ and $T_2 = (X_1 + X_2 + X_3)/3$, where X_1, X_2, X_3 is a random sample from a $N(\mu, \sigma)$ distribution, prove that T_2 is more efficient than T_1 .
- Given a random sample of size $n + 1$ from a $N(\mu, \sigma)$ distribution, show that the median, m , is roughly 64% less efficient than the sample mean for estimating the population mean. (Hint: In large samples $\text{Var}(m) = \pi\sigma^2/4n$.)
- Let X be a $\text{Bin}(n, \pi)$ random variable.

- Find the mean squared error of the π parameter estimators $T_1 = X/n$ and $T_2 = (X + 1)/(n + 2)$.
- When $n = 100$ and $\pi = 0.4$, which estimator, T_1 or T_2 , has the smaller *MSE*?
- Plot the efficiency of T_1 relative to T_2 versus π values in $(0, 1)$ for n values from 1 to 10.

- Given a random sample of size n from a $\Gamma(2, \lambda)$ distribution, consider the following estimators for $1/\lambda$:

$$T_1 = \frac{\bar{X}}{2} \quad \text{and} \quad T_2 = \frac{\sum_{i=1}^n X_i}{2(n+1)}$$

- Graph the relative efficiency of T_2 with regard to T_1 for values of λ from 0.01 to 100 with a sample size of 50.
 - Interpret the graph in (a).
 - Plot the relative efficiency of both estimators versus sample sizes from 1 to 100.
 - Interpret the graph in (c).
 - Generalize your findings.
- (Hint: $X \sim \Gamma(\alpha, \lambda)$, $E[X] = \alpha/\lambda$, $\text{Var}[X] = \alpha/\lambda^2$.)

- Consider a random variable $X \sim \text{Exp}(\lambda)$ and two estimators of $1/\lambda$, the expected value of X :

$$T_1 = \bar{X} \quad \text{and} \quad T_2 = \frac{\sum_{i=1}^n X_i + 1}{n+2}.$$

- Derive an expression for the relative efficiency of T_2 with respect to T_1 .
 - Plot $\text{eff}(T_2, T_1)$ versus n values of 1, 2, 3, 4, 20, 25, 30.
 - Generalize your findings.
- A baseball pitching machine launches fast balls whose speed follows a $N(\mu, \sigma = 5 \text{ km/h})$ distribution. Given the independent random samples \mathbf{X} and \mathbf{Y} , where $n_X = n_Y = 6$,

- (a) Show that the estimators $T_1 = \bar{X}$ and $T_2 = \frac{\sum_{i=2}^6 Y_i}{5}$ are unbiased estimators of μ .
- (b) Given the estimator $T_3 = \theta T_1 + (1 - \theta)T_2$, find the value of θ so that the MSE is a minimum.
8. Verify that $Var \left[\frac{\partial \ln f(X|\theta)}{\partial \theta} \right] = E \left[\left(\frac{\partial \ln f(X|\theta)}{\partial \theta} \right)^2 \right]$. (Hint: show that $E \left[\left(\frac{\partial \ln f(X|\theta)}{\partial \theta} \right) \right] = 0$.)
9. Verify that $E \left[\left(\frac{\partial \ln f(X|\theta)}{\partial \theta} \right)^2 \right] = -E \left[\left(\frac{\partial^2 \ln f(X|\theta)}{\partial \theta^2} \right) \right]$. (Hint: differentiate with respect to θ the equation $\int_{-\infty}^{\infty} \frac{\partial \ln f(x|\theta)}{\partial \theta} f(x|\theta) dx = 0$.)
10. The probability of obtaining a tail when flipping a coin can be $\pi = 1/2$, $\pi = 1/3$, or $\pi = 2/3$. To estimate π , the coin is flipped three times and one head is obtained on the first flip and tails on the second and third flips. Find the maximum likelihood estimator of π .
11. A manufacturer produces needles for a sewing machine in 5 units per parcel. The parcels are in boxes of 120 units. The manufacturer guarantees that only one out of 100 parcels is defective; however, the owner of a store thinks that at least 4 parcels out of 100 are defective. To solve the controversy, the manufacturer randomly chooses 18 boxes and checks the number of defective parcels. The results follow:

Number of defective parcels: 3, 1, 1, 2, 4, 2, 0, 1, 4, 1, 6, 2, 2, 3, 1, 4, 4, 2

Who is more likely to be right, the manufacturer or the store owner?

12. Given a random sample of size n from a geometric distribution,
- (a) Find the method of moments estimator of π .
- (b) Find the maximum likelihood estimator of π .
- (c) Use the results from (a) and (b) to compute the method of moments and maximum likelihood estimates from the sample $\{8, 1, 2, 0, 0, 0, 2, 1, 3, 3\}$, which represents the number of Bernoulli trials that resulted in failure before the first success in 10 experiments.
13. The following random samples $\mathbf{X}=(x_1, \dots, x_7)$ and $\mathbf{Y}=(y_1, \dots, y_{10})$ are drawn from $Pois(\lambda)$ and $Pois(2\lambda)$, respectively:

$\mathbf{X} \sim Pois(\lambda)$	4	2	5	7	3	4	3			
$\mathbf{Y} \sim Pois(2\lambda)$	6	10	1	6	3	5	5	4	7	5

- (a) Derive the maximum likelihood estimator of λ and calculate its variance.
- (b) Compute the maximum likelihood estimate of λ and its variance using the two random samples given.
14. Find the maximum likelihood estimator for μ if samples of size n are taken from a $N(\mu, \sigma = \sqrt{\mu})$ distribution.

- (a) Use the maximum likelihood estimator to calculate the maximum likelihood estimate that results from the sample

4.37, 9.30, 1.67, 1.25, 4.30, 6.97, 2.68, 5.49, 4.36, 4.46.

- (b) Plot the log-likelihood function versus μ for values between 4 and 5.

15. Given a random sample of size n from a distribution with a density function given by

$$f(x) = \theta \left(\frac{1}{x}\right)^{\theta+1}, \quad x \geq 1, \theta > 1,$$

- (a) Find the method of moments and the maximum likelihood estimators of θ .
 (b) Find the method of moments and maximum likelihood estimates of θ for the sample $\{3, 4, 2, 1.5, 4, 2, 3, 2, 4, 2\}$.
 (c) Set the seed equal to 42, and generate 1000 values from $f(x)$ using $\theta = 3$. Compute the method of moments and maximum likelihood estimates of θ using the generated values.

16. Given the density function

$$f(x) = (\theta + 1)(1 - x)^\theta, \quad 0 \leq x \leq 1, \theta > 0,$$

- (a) Find the maximum likelihood estimator of θ for a random sample of size n .
 (b) Set the seed equal to 88, and generate 1000 values from $f(x)$ when $\theta = 2$. Calculate the maximum likelihood estimate of θ from the generated values.
 (c) How close is the maximum likelihood estimate in (b) to $\theta = 2$?

17. Given the density function

$$f(x) = \frac{3}{\lambda} x^2 e^{-x^3/\lambda}, \quad x > 0, \quad \lambda > 0,$$

- (a) Find the maximum likelihood estimator of λ for a random sample of size n .
 (b) Verify that the maximum likelihood estimator is unbiased, consistent, and efficient.
 (c) Find the method of moments estimator of λ for a random sample of size n .

18. Given an exponential distribution with mean θ and the following estimators of θ :

$$\hat{\theta}_1 = X_1, \quad \hat{\theta}_2 = \frac{X_1 + X_2}{2}, \quad \hat{\theta}_3 = \bar{X}, \quad \hat{\theta}_4 = \min\{X_1, X_2, X_3\},$$

- (a) Find the mean and variance of each estimator.
 (b) Are any of the estimators efficient?
 (c) Which estimator is the MLE?
 (d) Let X be an exponential random variable with mean $\theta + 2$. Which estimator is an unbiased estimator of θ ?

19. Given a random sample of size n from a population of size N , where the items in the population are sequentially numbered from 1 to N ,

- (a) Derive the method of moments estimator of N .

- (b) Derive the maximum likelihood estimator of N .
- (c) What are the method of moments and maximum likelihood estimates of N for this sample of size 7: $\{2, 5, 13, 6, 15, 9, 21\}$?
20. The lifetime of a particular resistor follows an exponential distribution with parameter λ . The manufacturer claims the mean life of the resistor is 6 years. A distributor of the resistor is suing the manufacturer for excess warranty claims, saying that the mean life of the resistor is a mere 4 years. To resolve the issue, an accelerated test of the predicted lifetimes of 20 resistors is undertaken, yielding the following values:

3.70	1.76	3.63	15.73	5.85	0.20	9.87	14.55	0.43	2.46
0.45	5.09	10.53	12.41	3.19	3.41	3.80	1.66	0.40	1.10

- (a) The judge calls you as an expert witness to determine the validity of the suit. What do you tell the judge?
- (b) What value of λ maximizes the probability for values reported from the experiment.
- (c) Graph the log-likelihood function versus λ values ranging from 0 to 0.5.
21. Data frame `birthwt` from the `MASS` package has 10 variables recorded for each of 189 babies born at a U.S. hospital. The variable `low` takes the value 1 when the baby weighs less than 2.5 kg and 0 otherwise.
- (a) What distribution would be appropriate to model the values in `low`?
- (b) How many babies had birth weights less than 2.5 kg?
- (c) Find the maximum likelihood estimate for the parameter of the distribution selected in (a).
- (d) Interpret the MLE found in part (c).
22. In 1876, Charles Darwin had his book *The Effect of Cross- and Self-Fertilization in the Vegetable Kingdom* published. Darwin planted two seeds, one obtained by cross-fertilization and the other by auto-fertilization, in two opposite but separate locations of a pot. Self-fertilization, also called autogamy or selfing, is the fertilization of a plant with its own pollen. Cross-fertilization, or allogamy, is the fertilization with pollen of another plant, usually of the same species. Darwin recorded the plants' heights in inches. The data frame `Fertilize` from the `PASWR` package contains the data from this experiment.

Cross-fert	23.5	12.0	21.0	22.0	19.1	21.5	22.1	20.4
	18.3	21.6	23.3	21.0	22.1	23.0	12.0	
Self-fert	17.4	20.4	20.0	20.0	18.4	18.6	18.6	15.3
	16.5	18.0	16.3	18.0	12.8	15.5	18.0	

- (a) Create a variable `DD` defined as the difference between the variables `Cross-fert` and `Self-fert`.
- (b) Perform an exploratory analysis of `DD` to see if `DD` might follow a normal distribution.
- (c) Use the function `fitdistr()` found in the `MASS` package to obtain the maximum likelihood estimates of μ and σ if `DD` did follow a normal distribution.
- (d) Verify that the results from (c) are the sample mean and the uncorrected sample standard deviation of `DD`.

23. The lognormal distribution has the following density function:

$$g(w) = \frac{1}{w\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(\ln w - \mu)^2}, \quad w \geq 0, \quad -\infty < \mu < \infty, \quad \sigma > 0$$

where $\ln(w) \sim N(\mu, \sigma)$. The mean and variance of W are, respectively,

$$E[W] = e^{\mu + \frac{\sigma^2}{2}} \quad \text{and} \quad \text{Var}[W] = e^{2\mu + \sigma^2}(e^{\sigma^2} - 1).$$

Find the maximum likelihood estimators for $E[W]$ and $\text{Var}[W]$.

24. Consider the variable **brain** from the **Animals** data frame in the **MASS** package.

- Estimate with maximum likelihood techniques the mean and variance of **brain**. Specifically, use the R function `fitdistr()` with a lognormal distribution.
- Suppose that **brain** is a lognormal variable; then the log of this variable is normal. To check this assertion, plot the cumulative distribution function of **brain** versus a lognormal cumulative distribution function. In another plot, represent the cumulative distribution function of **log-brain** versus a normal cumulative distribution function. Is it reasonable to assume that **brain** follows a lognormal distribution?
- Find the mean and standard deviation of **brain** assuming a lognormal distribution.
- Repeat this exercise without the dinosaurs. Comment on the changes in the mean and variance estimates.

25. The data in **GD** available in the **PASWR** package are the times until failure in hours for a particular electronic component subjected to an accelerated stress test.

- Find the method of moments estimates of α and λ if the data come from a $\Gamma(\alpha, \lambda)$ distribution.
- Create a density histogram of times until failure. Superimpose a gamma distribution using the estimates from part (a) over the density histogram.
- Find the maximum likelihood estimates of α and λ if the data come from a $\Gamma(\alpha, \lambda)$ distribution by using the function `fitdistr()` from the **MASS** package.
- Create a density histogram of times until failure. Superimpose a gamma distribution using the estimates from part (c) over the density histogram.
- Plot the cumulative distribution for time until failure using the `ecdf()` function. Superimpose the theoretical cumulative gamma distribution using both the method of moments and the maximum likelihood estimates of α and λ . Which estimates appear to model the data better?

26. The time a client waits to be served by the mortgage specialist at a bank has density function

$$f(x) = \frac{1}{2\theta^3} x^2 e^{-x/\theta} \quad x > 0, \theta > 0.$$

- Derive the maximum likelihood estimator of θ for a random sample of size n .
- Show that the estimator derived in (a) is unbiased and efficient.
- Derive the method of moments estimator of θ .
- If the waiting times of 15 clients are 6, 12, 15, 14, 12, 10, 8, 9, 10, 9, 8, 7, 10, 7, and 3 minutes, compute the maximum likelihood estimate of θ .

27. If the function

$$f(x; \theta) = kx^3 e^{-\frac{1}{\theta}x} \quad x \geq 0, \theta > 0$$

is a density function,

- Find k .
- Derive the maximum likelihood estimator of θ for a random sample of size n .
- Derive the method of moments estimator of θ for a random sample of size n .
- Show that the estimators from parts (b) and (c) are both unbiased and efficient.

28. Given the function

$$f(x) = \frac{\theta}{x^2}, \quad x \geq \theta, \quad \theta > 0,$$

- Verify that it is a density function.
 - Find the maximum likelihood estimators of θ and $1/\theta$ for random samples of size n .
 - Is the maximum likelihood estimator of θ unbiased?
 - Find the method of moments estimators of θ and $1/\theta$.
29. The lifetime (in days) of a new 100 watt fluorescent light bulb follows an exponential distribution with mean λ . The following data are the lifetimes of 109 light bulbs:

Time	bubbles
[0, 50)	25
[50, 100)	19
[100, 150)	11
[150, 200)	8
[200, 250)	9
[250, 300)	7
[300, 450)	22
[450, 1050)	8

- Find the maximum likelihood estimator of λ .
 - Graph the logarithm of the likelihood function versus the parameter λ and indicate the value of λ where the lifetime is maximized.
30. Given the density function

$$f(x) = \frac{1}{\theta} x^{\frac{1-\theta}{\theta}}, \quad 0 < x < 1, \quad 0 < \theta < \infty$$

- Derive the maximum likelihood estimator of θ for a random sample of size n .
 - Derive the method of moments estimator of θ for a random sample of size n .
 - Show that the maximum likelihood estimator is unbiased.
31. Given the density function

$$f(x) = \theta x^{\theta-1} \quad 0 \leq x \leq 1, \quad \theta > 0$$

- (a) What distribution has this density function? Be sure to specify the parameter.
- (b) Find the maximum likelihood estimator of θ for random samples of size n .
- (c) Find the asymptotic variance of the maximum likelihood estimator.
- (d) Find the method of moments estimator of θ for a random sample of size n .
- (e) Calculate the maximum likelihood and method of moments estimates of θ for the sample $\{0.1, 0.7, 0.5, 0.85, 0.9\}$.

32. Given the density function

$$f(x) = \theta \left(\frac{1}{x}\right)^{\theta+1}, \quad x \geq 1, \theta > 1$$

- (a) Find the maximum likelihood estimator of θ for a random sample of size n .
- (b) Find the method of moments estimator of θ for a random sample of size n .
- (c) Calculate the maximum likelihood and method of moments estimates of θ using the sample values $\{2, 3, 2, 2.5, 1, 2, 2, 3, 1, 4, 6, 3, 4.4\}$.
- (d) Find the mean of the distribution.
- (e) Estimate the mean of the distribution using the maximum likelihood estimate of θ .

33. Given the density function

$$f(x) = \begin{cases} 1 - \theta & \text{for } -\frac{1}{2} \leq x \leq 0 \\ 1 + \theta & \text{for } 0 < x \leq \frac{1}{2} \end{cases}$$

- (a) Find the maximum likelihood estimator of θ for a random sample of size n .
 - (b) Show that the maximum likelihood estimator is unbiased and efficient.
- (Hint: Denote the number of observations as n_1 in the sample so that $0 < x_i \leq 1/2$.)

34. Given the density function

$$f(x) = 3\pi\theta x^2 e^{-\theta\pi x^3}, \quad x \geq 0.$$

- (a) Set the seed equal to 201, and generate a random sample of size $n = 500$ with $\theta = 5$.
- (b) Find the sample mean and the sample variance of the random values generated in (a).
- (c) Graph the density function.
- (d) Find the maximum likelihood estimate of θ .
- (e) Plot the logarithm of the likelihood function versus θ .

35. Given the density function

$$f(x) = e^{-(x-\alpha)}, \quad -\infty < \alpha \leq x.$$

- (a) Find the maximum likelihood and method of moments estimators of α .
- (b) Are both estimators found in (a) asymptotically unbiased?

36. Set the seed equal to 384, and generate 100 values from a $\beta(\alpha = 3, \beta = 2, A = 0, B = 1)$ distribution. Assume that these values are a random sample of size 100 from a β distribution with unknown parameters. Use maximum likelihood techniques to obtain estimates of α and β from this sample.
37. Given a random sample of size n from an exponential distribution with **pdf**

$$f(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}} \quad x \geq 0, \quad \theta > 0, \quad (7.68)$$

- (a) Find the MLE of θ .
- (b) Given the answer in part (a), what is the MLE of θ^2 .

Chapter 8

Confidence Intervals

8.1 Introduction

In Chapter 7, techniques to find point estimators, such as the method of moments and maximum likelihood, were introduced as well as were criteria to evaluate the “goodness” of an estimator. However, even the most efficient unbiased estimator is not likely to estimate the population parameter exactly. Further, a point estimate provides no information about the precision or reliability of the estimate. Consequently, the construction of an **interval estimate** or **confidence interval (CI)**, where the user can control the precision (width) of the interval as well as the reliability (confidence) that the true parameter will be found in the confidence interval, is desirable.

A $(1 - \alpha)$ confidence interval for a parameter θ , denoted $CI_{1-\alpha}(\theta)$, is constructed by first selecting a confidence level, denoted by $(1 - \alpha)$ and typically expressed as a percentage, $(1 - \alpha) \cdot 100\%$. The confidence level is simply a measure of the degree of reliability in the procedure used to construct the confidence interval. Typical confidence levels are 90%, 95%, or 99%. A confidence level of 99% implies that 99% of all samples would provide confidence intervals that would contain θ . Clearly, it is desirable to have a high degree of reliability. However, with increased reliability, the width of the confidence interval increases. So, the goal is to construct a confidence interval with a width the practitioner finds useful while maintaining a degree of reliability that is as high as possible. The relationship between the width and confidence level in a confidence interval will become clearer once some actual confidence interval formulas are examined. The confidence interval has two limits, a lower limit denoted by $L(\mathbf{X})$ and an upper limit denoted by $U(\mathbf{X})$. The **confidence level** is defined as $\mathbb{P}(\theta \in [L(\mathbf{X}), U(\mathbf{X})])$. That is, an interval should be constructed such that

$$\mathbb{P}(L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})) = 1 - \alpha. \quad (8.1)$$

It is important to note that the interval $[L(\mathbf{X}), U(\mathbf{X})]$ is a random interval since it depends on the random variables of \mathbf{X} . However, after a sample is obtained and values for $[L(\mathbf{X}), U(\mathbf{X})]$ are calculated, the probability that the parameter θ will be included in the interval $[L(\mathbf{x}), U(\mathbf{x})]$ is either 0 or 1, depending, of course, on whether θ is between the lower limit $L(\mathbf{x})$ and the upper limit $U(\mathbf{x})$. Note that \mathbf{X} changes to an \mathbf{x} once there are values, x_i , for the random variables, X_i . The probability the parameter θ is contained in the random interval $[L(\mathbf{X}), U(\mathbf{X})]$ from (8.1) is $(1 - \alpha)$. However, once the values for the random variables are observed, (8.1) is written as

$$CI_{1-\alpha}(\theta) = [L(\mathbf{x}), U(\mathbf{x})], \quad (8.2)$$

which is called a $(1 - \alpha)$ confidence interval. Consequently, it makes no sense to talk about a $(1 - \alpha)$ probability interval. Frequently, the confidence level is expressed as a percentage, and the interval is often called a $(1 - \alpha) \cdot 100\%$ confidence interval. A $(1 - \alpha) \cdot 100\%$ confidence interval is typically interpreted as “One is $(1 - \alpha) \cdot 100\%$ confident θ is contained

in the interval $[L(\mathbf{x}), U(\mathbf{x})]$.” However, the word *confidence* in such statements applies to the procedure used to construct the interval, not the interval itself. That is, if there were an infinite number of samples, $(1 - \alpha) \cdot 100\%$ of them would contain θ .

Confidence intervals of the form $[L(\mathbf{x}), U(\mathbf{x})]$ are referred to as two-sided confidence intervals. However, some applications will only require a single bound. For example, only a lower confidence bound on the mean shear strength of an aluminum tube is required to ensure the minimum design specification for a top tube of a bicycle is met. Likewise, only an upper confidence bound on the mean level of NO_3 in potable water is required to ensure the maximum allowable limit is not exceeded. One-sided confidence intervals take the form

$$\mathbb{P}(L(\mathbf{X}) \leq \theta) = 1 - \alpha \quad \text{or} \quad \mathbb{P}(\theta \leq U(\mathbf{X})) = 1 - \alpha,$$

depending on whether the confidence interval is a lower confidence interval, $[L(\mathbf{x}), \infty)$, or an upper confidence interval, $(-\infty, L(\mathbf{x})]$, respectively. Unless otherwise specified, a confidence interval will refer to a two-sided confidence interval.

There are several techniques used to obtain both one-sided and two-sided confidence intervals. One of the more popular methods for constructing confidence intervals uses pivotal quantities. A random variable $Q(\mathbf{X}; \theta)$ is a **pivotal quantity** or **pivot** if the distribution of Q is independent of the parameter θ . A method of constructing confidence intervals using pivots is introduced in Section 8.2.1 and is used to derive most of the confidence interval formulas in this chapter.

8.2 Confidence Intervals for Population Means

8.2.1 Confidence Interval for the Population Mean when Sampling from a Normal Distribution with Known Population Variance

A random sample of size n is taken from a normal distribution with mean μ and variance σ^2 . To obtain a confidence interval for μ , recall that the sampling distribution for the sample mean is $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$. Using the sampling distribution of \bar{X} , create the pivotal quantity

$$Q(\mathbf{X}; \mu) = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1). \quad (8.3)$$

To obtain a confidence interval with a $(1 - \alpha)$ confidence level, construct a region such that the area between $z_{\alpha/2}$ and $z_{1-\alpha/2}$ is $(1 - \alpha)$, as shown in Figure 8.1 on the next page. In other words,

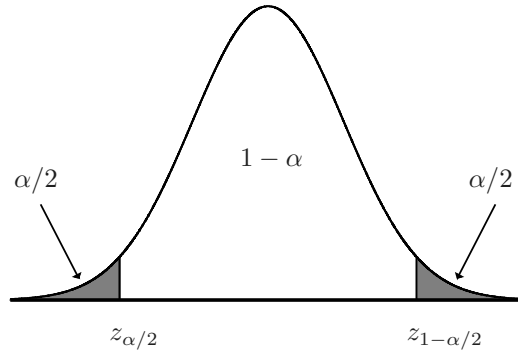
$$\mathbb{P}\left(z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}\right) = 1 - \alpha. \quad (8.4)$$

Multiply both sides of (8.4) by σ/\sqrt{n} , to obtain

$$\mathbb{P}\left(z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Subtract \bar{X} from both sides, multiply both sides by -1 , and rearrange the inequalities, to get

$$\mathbb{P}\left(\bar{X} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

FIGURE 8.1: Standard normal distribution with an area of $\alpha/2$ in each tail

Consequently, the $(1-\alpha)$ confidence interval for μ , when sampling from a normal distribution with known variance, is given by

$$\left[\bar{x} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right],$$

or, equivalently, by recognizing that $z_{\alpha/2} = -z_{1-\alpha/2}$, write the standard form as

$$\boxed{CI_{1-\alpha}(\mu) = \left[\bar{x} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right].} \quad (8.5)$$

Note that \bar{X} in the probability statement changes to \bar{x} in the confidence interval formula.

To obtain a one-sided (either upper or lower) confidence interval in a symmetric distribution, proceed in a similar fashion. That is, write

$$\mathbb{P}\left(-z_{1-\alpha} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right) = 1 - \alpha \quad \text{or} \quad \mathbb{P}\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha}\right) = 1 - \alpha$$

and rearrange the quantities inside the probability statements to obtain

$$\mathbb{P}\left(\mu \leq \bar{X} + z_{1-\alpha} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \quad \text{or} \quad \mathbb{P}\left(\bar{X} - z_{1-\alpha} \frac{\sigma}{\sqrt{n}} \leq \mu\right) = 1 - \alpha$$

Thus,

$$UCI_{1-\alpha}(\mu) = \left(-\infty, \bar{x} + z_{1-\alpha} \frac{\sigma}{\sqrt{n}}\right) \quad \text{or} \quad LCI_{1-\alpha} = \left[\bar{x} - z_{1-\alpha} \frac{\sigma}{\sqrt{n}}, \infty\right)$$

Note that a one-sided confidence interval can be obtained from a two-sided confidence interval by simply changing the $z_{1-\alpha/2}$ value to a $z_{1-\alpha}$ value and replacing the lower or upper limit with $-\infty$ or ∞ , respectively, depending on whether an upper or lower confidence interval is desired.

Example 8.1 Generate 100 samples, each of size 500, from a $N(0, 1)$ distribution. For each of the 100 samples of size 500, calculate a 95% confidence interval for the population mean. Finally, determine how many of the 100 intervals contain the population mean, $\mu = 0$. This number is the simulated confidence level.

Solution: The function `interval.plot()` graphically depicts the confidence intervals that are simulated:

```
> interval.plot<- function(ll, ul){
+   y1 <- ll ; y2<-ul; n <- length(y1)
+   plot(y1, type = "n", ylim=c(-.3,.3), xlab = " ", ylab = " ")
+   condition <- (ll <= 0 & ul >= 0)
+   segments((1:n)[y1<0&y2>0], y1[y1<0&y2>0],(1:n)[y1<0&y2>0],
+   y2[y1<0&y2>0])
+   segments((1:n)[y1>0], y1[y1>0],(1:n)[y1>0], y2[y1>0], col=17,
+   lwd=8)
+   segments((1:n)[y2<0], y1[y2<0],(1:n)[y2<0], y2[y2<0], col=17,
+   lwd=8)
+   SUM<-sum(condition)
+   abline(h=0)
+   cat("Number of intervals that contain 0 =", SUM,"\n" )}
```

Next, write a script that calculates the lower and upper limits of the confidence intervals. The lower limit is indicated by `ll` and the upper limit is indicated by `ul` in the following script:

```
> set.seed(402)
> m<-100 # Number of samples
> n<-500 # Sample size
> a<-array(0, m)
> ll<-array(0, m)
> ul<-array(0, m)
> i<-0
> while (i<m) {i<-i+1
+ a[i] <-mean(rnorm(n))
+ ll[i] <-a[i]+qnorm(0.025)*sqrt(1/n)
+ ul[i] <-a[i]+qnorm(0.975)*sqrt(1/n)}
> interval.plot(ll, ul)
Number of intervals that contain 0 = 95
```

Note that this is a random simulation and consequently the number of confidence intervals that contain zero will vary and will not always equal the expected 95. A graphical representation of confidence intervals using a different seed value with the function `interval.plot()` is found in Figure 8.2 on the facing page. A more general function that can be used to generate random data and subsequently to create confidence intervals is the function `CIsim()` from the PASWR package. ■

Example 8.2 A random sample of size 30 is taken from a normal distribution with unknown mean μ and standard deviation $\sigma = 2.5$. Given that $\sum_{i=1}^{30} x_i = 77$, calculate a 95% confidence interval for the population mean.

Solution: First, determine \bar{x} :

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{77}{30} = 2.57$$

Since the sample was taken from a normal distribution with known variance, it is permissible to write

$$\mathbb{P} \left(\bar{X} - z_{0.975} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{0.975} \frac{\sigma}{\sqrt{n}} \right) = 0.95.$$

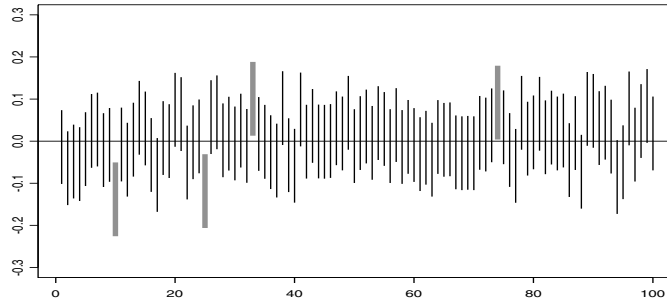


FIGURE 8.2: Simulated confidence intervals for the population mean when sampling from a normal distribution with known variance

A 95% confidence interval for the mean using (8.5) is written as

$$CI_{0.95}(\mu) = \left[2.57 - (1.96) \frac{2.5}{\sqrt{30}}, 2.57 + (1.96) \frac{2.5}{\sqrt{30}} \right].$$

In other words, one can be 95% confident that the mean, μ , will be found in the interval $CI_{0.95}(\mu) = [1.68, 3.46]$. It is important to note that the sample mean ($\bar{x} = 2.57$) is the center point of this interval; however, this will only be the case in symmetric distributions. ■

Example 8.3 ▷ **Confidence Interval for μ : Grocery Spending** ◁ The consumer expenditure survey, created by the U.S. Department of Labor, was administered to 30 households in Watauga County, North Carolina, to see how the cost of living in Watauga County with respect to total dollars spent on groceries compares with other counties. The amount of money each household spent per week on groceries is given in Table 8.1 and stored in the data frame **Grocery**.

- Construct a 97% confidence interval for the true mean weekly grocery expenditure for Watauga County households. Historical records indicate that the variance for grocery expenditure per household in Watauga County is 900 dollars².
- A grocery chain is considering building a new grocery store in Watauga County. However, it will only do so if it is 99% confident the average amount spent on groceries each week is at least \$105. Does a $LCI_{0.99}(\mu)$ include \$105? If so, what does that imply?

Table 8.1: Weekly spending in dollars (**Grocery**)

90.74	104.02	85.64	134.71	108.85	142.19	162.87	138.2	98.73	98.18
139.84	159.69	147.03	151.16	105.68	116.93	97.46	146.64	90.92	134.54
110.82	109.90	106.74	122.10	152.28	136.01	126.00	108.69	135.06	57.38

Solution: The answers are as follows:

(a) Before using (8.5), the confidence interval formula for μ with known σ on page 293, it is necessary to verify that the assumption of normality is satisfied. To do this, create a normal quantile-quantile plot using the `qqnorm()` function as follows:

```
> attach(Grocery)
> qqnorm(groceries)
> qqline(groceries)
```

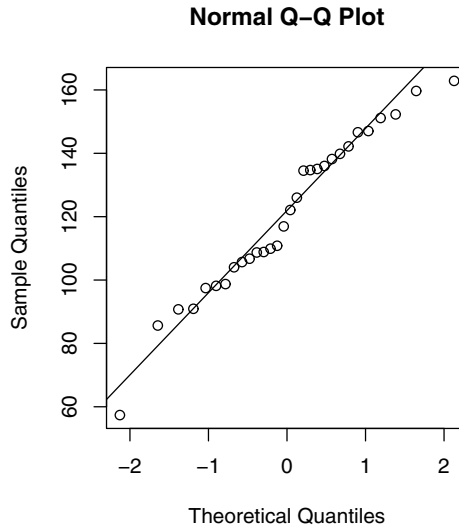


FIGURE 8.3: Quantile-quantile (normal distribution) plot of weekly monies spent on groceries for 30 randomly selected Watauga households

The resulting normal quantile-quantile plot is shown in Figure 8.3. Note that the plotted values fall relatively close to the plotted line, indicating the assumption of normality is reasonable. Consequently, one decides the assumption for using (8.5) on page 293 is satisfied and continues by finding the sample mean:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{3619}{30} = 120.63$$

Using the historical value of 900 for σ^2 , the 97% confidence interval is given by

$$\begin{aligned} CI_{0.97}(\mu) &= \left[\bar{x} - z_{0.985} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{0.985} \frac{\sigma}{\sqrt{n}} \right] \\ &= \left[120.63 - (2.17) \frac{\sqrt{900}}{\sqrt{30}}, 120.63 + (2.17) \frac{\sqrt{900}}{\sqrt{30}} \right]. \end{aligned}$$

In other words, one can be 97% confident that the mean grocery spending will be found in the interval [108.75, 132.52] dollars.

To do this calculation with S, enter

```
> mean(groceries)
[1] 120.6333
> qnorm(0.03/2)
[1] -2.17009
> round(c(mean(groceries)-qnorm(1-0.03/2)*sqrt(900/30),
+         mean(groceries)+qnorm(1-0.03/2)*sqrt(900/30)), 2)
[1] 108.75 132.52 #97% CI
```

(b) Part (a) already verified that the data follow a normal distribution, so one calculates the one-sided 99% confidence interval as

$$\begin{aligned} LCI_{0.99}(\mu) &= \left[\bar{x} - z_{0.99} \frac{\sigma}{\sqrt{n}}, \infty \right) \\ &= \left[120.63 - 2.33 \frac{30}{\sqrt{30}}, \infty \right) \\ &= [107.87, \infty) \end{aligned}$$

This interval does not include \$105; however, its lower limit is above \$105, so the grocery chain can be more than 99% confident the mean grocery spending is greater than \$105. ■

8.2.1.1 Determining Required Sample Size

Larger sample sizes generally result in narrower confidence intervals. Researchers will often desire a confidence interval not to exceed a specific width at some predetermined level of significance (one that usually has some practical significance to their research). The problem addressed in this section is how to determine the minimum required sample size to be within a given distance of μ when estimating the population mean with known variance, σ^2 . To start, recall the probability statement about μ for normal distributions with known variance in (8.6). Use this equation when working with normal populations ($N(\mu, \sigma)$), as well as with various other distributions, provided the sample size is sufficiently large:

$$\mathbb{P} \left(\bar{X} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right) = 1 - \alpha, \quad (8.6)$$

which implies

$$\mathbb{P} \left(|\bar{X} - \mu| \leq z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right) = 1 - \alpha,$$

where $|\bar{X} - \mu|$ is the **error of estimation**. In general, the error of estimation is a measure of the goodness of the estimate. Many texts refer to the error of estimation as the **margin of error** or the **bound on the error**. Denote this quantity by B . If one assumes the maximum error is

$$B = |\bar{x} - \mu| = z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}},$$

one can solve for n as shown in (8.7):

$$\boxed{n = \frac{(z_{1-\alpha/2})^2 \sigma^2}{(\bar{x} - \mu)^2} = \left(\frac{z_{1-\alpha/2} \sigma}{B} \right)^2} \quad (8.7)$$

Consequently, any time a confidence level is specified and the value of σ is known, one can determine the required sample size, n , to be within the maximum error, B , that is acceptable.

Example 8.4 Determine the required sample size to estimate the true value of μ within ± 0.02 with a confidence level of 95% when sampling from a normal distribution with $\sigma = 0.1$.

Solution: To determine the required sample size, use (8.7) as follows:

$$n = \frac{(z_{1-\alpha/2})^2 \sigma^2}{(\bar{x} - \mu)^2} = \frac{(1.96)^2 (0.1)^2}{(0.02)^2} = 96.04.$$

In order to have a confidence of at least $1 - \alpha$, one always takes the ceiling of n . Therefore, the required sample size n to estimate the population mean with a 0.95 confidence level so that the margin of error is no more than 0.02 is $n = 97$. ■

Example 8.5 Suppose a random sample of size n from a normal distribution with unknown mean μ and standard deviation $\sigma = 5$ is taken. Calculate the minimum sample size so that one can be 95% confident the interval $[\bar{x} - 1, \bar{x} + 1]$ contains the true value of μ .

Solution: Given that the sample was taken from a normal distribution with known variance, one can write

$$\mathbb{P}\left(\bar{X} - (1.96)\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + (1.96)\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

Since one needs to be 95% confident the interval $[\bar{x} - 1, \bar{x} + 1]$ contains μ , write

$$\mathbb{P}(\bar{X} - 1 \leq \mu \leq \bar{X} + 1) = 0.95,$$

set $1.96 \frac{\sigma}{\sqrt{n}} = 1$, and solve for n given that $\sigma = 5$:

$$n = (1.96)^2 (5)^2 = 96.04$$

Since a sample of 96.04 is impossible, take the ceiling of n to make sure the confidence is at or above the specified level. Consequently, the minimum sample size to be at least 95% confident the interval $[\bar{x} - 1, \bar{x} + 1]$ contains μ is $n = 97$ when $\sigma = 5$. ■

Example 8.6 ▷ *Sample Size: Defective Containers* ◁ In a company that produces glass containers, the probability of producing a defective container is $\pi = 0.03$, and the probability of obtaining a functional container is $(1 - \pi) = 0.97$. Determine how many containers need to be manufactured to guarantee that at least 100 containers are defective with a probability of at least 0.95.

Solution: Three solutions are presented for this problem. The first is the exact answer based on a negative binomial distribution and requires the use of a computer. The second is an approximation that can be used in the absence of a computer. The third is the exact answer from a $(Bin(n, 0.03))$, the distribution approximated in (b).

- (a) Let X be the number of failures prior to the $r = 100^{\text{th}}$ success (defective container). The distribution of X is $NB(100, 0.03)$. The problem requests $\mathbb{P}(X = x | 100, 0.03) \geq 0.95$. That is, one must find the number x of non-defective containers to guarantee the probability is at least 0.95 upon obtaining the 100^{th} defective container. The following S code indicates the total number of containers that must be manufactured to guarantee 100 are defective with a probability over 0.95 is 3891:

```
> f <- 0      # f= number of failures
> p <- 0      # p = probability
```

```

> s <- 100    # s = number of successes
> while(p < 0.95){
+   f <- f + 1
+   p <- pbinom(f, s, 0.03)}
> ans <- c(f + s, p) # f+s= Containers
> names(ans) <- c("Containers", "Probability")
> ans
  Containers Probability
      3891    0.9500444

```

- (b) Let the random variable X represent the number of defective containers. The distribution of X is $Bin(n, 0.03)$. Consequently,

$$E[X] = n\pi = 0.03n \quad \text{and} \quad Var[X] = n\pi(1 - \pi) = 0.0291n.$$

If it is assumed n is sufficiently large and the production of each container is independent, one can approximate the distribution of X using a normal distribution where

$$\mathbb{P}(X \geq 100) = 0.95.$$

Equivalently,

$$\mathbb{P}\left(\frac{X - n\pi}{\sqrt{n\pi(1 - \pi)}} \geq \frac{100 - n\pi}{\sqrt{n\pi(1 - \pi)}}\right) = \mathbb{P}\left(Z \geq \frac{100 - 0.03n}{\sqrt{0.0291n}}\right) = 0.95.$$

Note that $\mathbb{P}(Z \geq -z_{1-\alpha}) = 0.95 \Rightarrow -z_{1-\alpha} = -z_{0.95} = -1.64$, and solve the equation

$$\frac{100 - 0.03n}{\sqrt{0.0291n}} = 1.64, \quad (8.8)$$

which is equivalent to solving

$$0.0009n^2 - 6.07826n + 100^2 = 0. \quad (8.9)$$

The solutions to (8.9) are $n \approx 3924$ or 2832 . However, the value 2832 is not acceptable since it does not satisfy (8.8). Consequently, the number of containers the factory needs to manufacture to be 95% confident of getting at least 100 defective containers is 3924.

- (c) Let the random variable X again represent the number of defective containers. The distribution of X is $Bin(n, 0.03)$. To solve $\mathbb{P}(X \geq 100) \geq 0.95$ with `S`, use code similar to what follows and keep in mind that $\mathbb{P}(X \geq 100) = 1 - \mathbb{P}(X \leq 99)$. The following `S` code indicates the total number of containers that must be manufactured to guarantee 100 are defective with a probability over 0.95 is 3891 when using the binomial random variable. This agrees with the answer that was found when modeling the number of defective containers obtained with a negative binomial random variable.

```

> n <- 0    # Number of containers
> p <- 0    # Probability
> while(p < 0.95) {
+   n <- n + 1
+   p <- 1 - pbinom(99, n, 0.03)}
> ans <- c(n, p)
> names(ans) <- c("Containers", "Probability")
> ans
  Containers Probability
      3891    0.9500444

```



The confidence intervals discussed in the remainder of this chapter are commonly used confidence intervals based, for the most part, on the normal distribution. When constructing confidence intervals, if historical evidence does not support normality or the text narrative does not explicitly specify the sample information was collected from a normal distribution, one should not blindly use techniques that require the normality assumption! Checking normality assumptions graphically with normal quantile-quantile plots as discussed in Section 4.3.7 on page 158 should become a habit.

8.2.2 Confidence Interval for the Population Mean when Sampling from a Normal Distribution with Unknown Population Variance

Suppose a random sample of size n is taken from a normal distribution with unknown mean μ and unknown variance σ^2 . To construct a confidence interval for μ , use the pivotal quantity

$$Q(\mathbf{X}; \mu) = \frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t_{n-1}.$$

Operating in a similar fashion to the derivation of the confidence interval for μ , using (8.3) from Section 8.2.1, one obtains the interval

$$CI_{1-\alpha}(\mu) = \left[\bar{x} - t_{1-\alpha/2; n-1} \frac{s}{\sqrt{n}}, \bar{x} + t_{1-\alpha/2; n-1} \frac{s}{\sqrt{n}} \right]. \quad (8.10)$$

Example 8.7 A random sample of size 12 is taken from a population that follows a $N(\mu, \sigma)$ distribution where the value for σ is unknown. Given:

$$\sum_{i=1}^{12} x_i = 61.9, \quad \text{and} \quad \sum_{i=1}^{12} x_i^2 = 450,$$

determine a 90% confidence interval for the population mean.

Solution: First find the sample mean and the sample variance.

$$\begin{aligned} \bar{x} &= \frac{\sum_{i=1}^n x_i}{n} = \frac{61.9}{12} = 5.16, \text{ and} \\ s^2 &= \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{n-1} = \frac{450 - (12)(5.16)^2}{12-1} = 11.86. \end{aligned}$$

The sample standard deviation is $s = 3.44$ and $t_{0.95; 11} = 1.8$. Using (8.10),

$$CI_{0.9}(\mu) = \left[5.16 - (1.8) \frac{3.44}{\sqrt{12}}, 5.16 + (1.8) \frac{3.44}{\sqrt{12}} \right] = [3.37, 6.95].$$

One is 90% confident the population mean lies in $[3.37, 6.95]$. The value $t_{0.95; 11}$ can be found by using the S command `qt(0.95, 11)`. ■

Example 8.8 ▷ *Confidence Interval for μ : House Prices* ◁ Estimate the mean house price for three-bedroom/two-bath houses in Watauga County, North Carolina. A random sample of 14 three-bedroom/two-bath houses was taken from the Watauga County Multiple Listing Service real estate listings (2003), and the results are reported in Table 8.2 on the next page and stored in the data frame `House`. Calculate a 95% confidence interval for the average price of a three bedroom/two bath house in this county.

Table 8.2: House prices (in thousands of dollars) for three-bedroom/two-bath houses in Watauga County, NC (**House**)

Neighborhood	Price	Neighborhood	Price
Valley Crucis	184.9	Blowing Rock	279.5
Valley Crucis	160.0	Valley Crucis	294.9
Valley Crucis	298.0	Blowing Rock	324.5
Blowing Rock	269.9	Blowing Rock	226.0
Parkway	189.9	Valley Crucis	329.9
Blowing Rock	229.9	Green Valley	199.9
Cove Creek	175.0	Park Valley	133.9

Solution: Before using the confidence interval formula in (8.10), one needs to verify the assumption of normality is satisfied. Consequently, a normal quantile-quantile plot for the values reported in Table 8.2 was constructed with the S functions `qqnorm()` and `qqline()` and is shown in Figure 8.4 on the next page. Since the points in Figure 8.4 fall relatively close to the straight line, it is decided that the normality assumption for using (8.10) is satisfied. Thus, continue by calculating the sample mean as

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{3296.2}{14} = 235.44$$

and the sample variance as

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} = \frac{\sum_{i=1}^{14} (x_i - 235.44)^2}{13} = 4084.4.$$

The sample standard deviation is $s = 63.91$, and a 95% confidence interval using (8.10) is calculated as

$$\begin{aligned} CI_{0.95}(\mu) &= \left[\bar{x} - t_{0.975;n-1} \frac{s}{\sqrt{n}}, \bar{x} + t_{0.975;n-1} \frac{s}{\sqrt{n}} \right] \\ &= \left[235.44 - (2.16) \frac{63.91}{\sqrt{14}}, 235.44 + (2.16) \frac{63.91}{\sqrt{14}} \right] = [198.54, 272.34]. \end{aligned}$$

Thus, one is 95% confident the mean house price falls in [198.54, 272.34] thousands of dollars.

To construct a confidence interval for the mean with S, type

```
> attach(House)
> MEAN<-mean(Price)
> CT<-qt(.975,13)
> ST<-sd(price) #stdev(Price) in S-PLUS
> round(c(MEAN-CT*ST/sqrt(14), MEAN+CT*ST/sqrt(14)),2)
[1] 198.54 272.34
```

Direct construction of the confidence interval is also possible using the internal function `t.test()` as shown next.

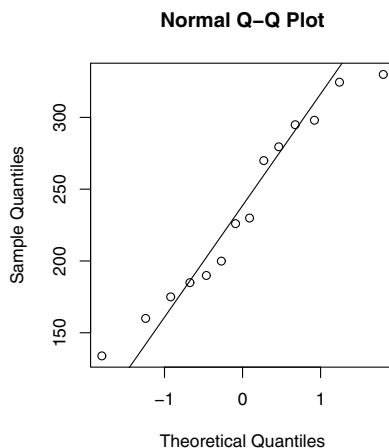


FIGURE 8.4: Quantile-quantile plot of the asking price for 14 randomly selected three-bedroom/two-bath houses in Watauga County, North Carolina

```
> t.test(Price)$conf
[1] 198.5424 272.3433
attr("conf.level")
[1] 0.95
```

Note that the function `sd(object)` finds the standard deviation in R but will not work with S-PLUS. The function to find the standard deviation with S-PLUS is `stdev(object)`. The default confidence level is 95% for both R and S-PLUS. To change the confidence level, say to 90%, the argument `conf.level=.90` is specified inside the `t.test()` command as `t.test(object, conf.level=.90)$conf`. ■

8.2.3 Confidence Interval for the Difference in Population Means when Sampling from Independent Normal Distributions with Known Equal Variances

Consider two normal and independent populations $N(\mu_X, \sigma_X)$ and $N(\mu_Y, \sigma_Y)$, where $\sigma_X = \sigma_Y = \sigma$ is known. If one takes random samples of sizes n_X and n_Y , respectively, a confidence interval for $\mu_X - \mu_Y$ is easily derived using the sampling distribution of

$$\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \sigma\sqrt{\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}\right),$$

which provides a pivotal quantity,

$$Q(\mathbf{X}, \mathbf{Y}; \mu_X - \mu_Y) = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sigma\sqrt{\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}},$$

which has a standard normal distribution independent of the value of $\mu_X - \mu_Y$. The $(1 - \alpha) \cdot 100\%$ confidence interval for the difference in population means, $\mu_X - \mu_Y$, is given by

$$\begin{aligned}
 & CI_{1-\alpha}(\mu_X - \mu_Y | \sigma_X = \sigma_Y \text{ is known}) = \\
 & \left[(\bar{x} - \bar{y}) - z_{1-\alpha/2}\sigma\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}, (\bar{x} - \bar{y}) + z_{1-\alpha/2}\sigma\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}} \right]. \quad (8.11)
 \end{aligned}$$

Example 8.9 Suppose independent random samples are taken from two normal distributions $N(\mu_X, \sigma = 3)$ and $N(\mu_Y, \sigma = 3)$, respectively, such that $n_X = 15$, $\sum_{i=1}^{n_X} x_i = 60$, $n_Y = 22$, and $\sum_{i=1}^{n_Y} y_i = 97$. Calculate a 95% confidence interval for the difference in population means ($\mu_X - \mu_Y$).

Solution: Since

$$\bar{x} = \frac{\sum_{i=1}^{n_X} x_i}{n_X} = \frac{60}{15} = 4 \quad \text{and} \quad \bar{y} = \frac{\sum_{i=1}^{n_Y} y_i}{n_Y} = \frac{97}{22} = 4.41,$$

the 95% confidence interval for the difference in population means ($\mu_X - \mu_Y$) is calculated using (8.11) as

$$\begin{aligned}
 CI_{0.95}(\mu_X - \mu_Y) &= \left[(4 - 4.41) - (1.96)(3)\sqrt{\frac{1}{15} + \frac{1}{22}}, (4 - 4.41) + (1.96)(3)\sqrt{\frac{1}{15} + \frac{1}{22}} \right] \\
 &= [-2.38, 1.56].
 \end{aligned}$$

To calculate the confidence interval with S, key in

```

> round(qnorm(0.975), 2)
[1] 1.96
> round(c((4-4.41) + qnorm(0.025)*3*sqrt(1/15 + 1/22),
+ (4-4.41) + qnorm(0.975)*3*sqrt(1/15 + 1/22)), 2)
[1] -2.38 1.56

```

So, one is 95% confident $\mu_X - \mu_Y$ lies in $[-2.38, 1.56]$. ■

Example 8.10 The hardness of a piece of fruit is a good indicator of the fruit's ripeness. An experiment was undertaken where 17 recently picked (fresh) apples were randomly selected and measured for hardness. Seventeen apples were also randomly selected from a warehouse where the apples had been stored for one week. Construct a 95% confidence interval for the mean difference between the hardness of fresh apples and the hardness for apples that were picked one week ago. Assume the distributions for both recently picked apples and for apples picked one week ago have known and equal variances of $2.25 \text{ (kg/meter}^2\text{)}^2$. The data are provided in Table 8.3 on the following page and can be found in the data frame **Apple**.

Solution: Before the confidence interval formula in (8.11) can be used, one needs to make sure the assumption of normality is satisfied. Consequently, a normal quantile-quantile plot for the values reported in Table 8.3 on the next page was constructed and is shown in Figure 8.5.

(Note that title and axis labels in Figure 8.5 are the R defaults and that a “Q-Q Normal Plot” is equivalent to a normal quantile-quantile plot as discussed earlier.) Since the points in Figure 8.5 fall relatively close to the straight lines, it is decided that the normality assumptions for using (8.11) are satisfied.

Table 8.3: Apple hardness measurements (Apple)

Fresh			Warehouse		
7.27	8.38	9.20	7.79	9.17	10.05
6.65	5.83	7.89	7.11	6.31	8.58
5.76	7.70	7.77	6.27	8.39	8.42
6.53	5.86	6.48	7.22	6.19	7.07
8.09	5.53	8.28	8.83	6.31	8.83
9.56	6.54		10.5	7.17	

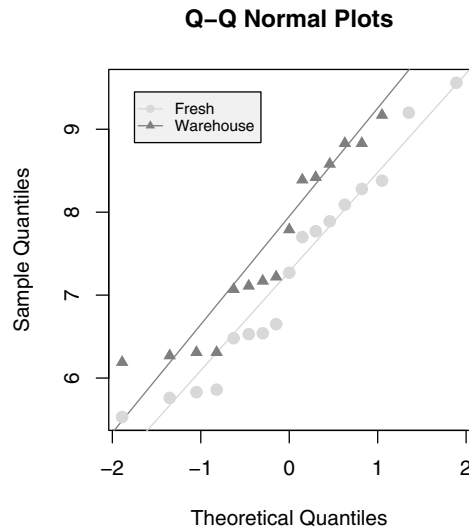


FIGURE 8.5: Superimposed normal quantile-quantile plots of the hardness values for fresh and warehoused apples

The R code used to produce Figure 8.5 is

```

> attach(Apple)
> par(pty = "s")
> Altblue <- "#A9E2FF"
> Adkblue <- "#0080FF"
> fresh <- qqnorm(Fresh)
> old <- qqnorm(Warehouse)
> plot(fresh,type="n",ylab="Sample Quantiles",xlab="Theoretical Quantiles")
> qqline(Fresh, col = Altblue)
> qqline(Warehouse, col = Adkblue)
> points(fresh, col = Altblue, pch = 16, cex = 1.2)
> points(old, col = Adkblue, pch = 17)
> legend(-1.75, 9.45, c("Fresh", "Warehouse"), col = c(Altblue, Adkblue),
+ text.col=c("black","black"),pch=c(16,17),lty=c(1,1),bg="gray95",cex=0.75)
> title("Q-Q Normal Plots")

```

Thus, continue solving the problem by calculating the sample mean hardness for both the fresh and warehoused apples as

$$\bar{x} = \frac{\sum_{i=1}^{n_X} x_i}{n_X} = \frac{123.25}{17} = 7.25 \quad \text{and} \quad \bar{y} = \frac{\sum_{i=1}^{n_Y} y_i}{n_Y} = \frac{134.13}{17} = 7.89.$$

Using (8.11), the 95% confidence interval for $\mu_X - \mu_Y$ is

$$CI_{0.95}(\mu_X - \mu_Y) = \left[(7.25 - 7.89) - (1.96)(1.5)\sqrt{\frac{1}{17} + \frac{1}{17}}, \right. \\ \left. (7.25 - 7.89) + (1.96)(1.5)\sqrt{\frac{1}{17} + \frac{1}{17}} \right] = [-1.65, 0.37].$$

To calculate the confidence interval with S, enter

```
> attach(Apple)
> str(Apple) # Only works in R
'data.frame': 17 obs. of 2 variables:
 $ Fresh : num 7.27 6.65 5.76 6.53 8.09 9.56 8.38 ...
 $ Warehouse: num 7.79 7.11 6.27 7.22 8.83 10.5 9.17 ...
> mean.fresh <- mean(Fresh)
> mean.fresh
[1] 7.254118
> mean.old <- mean(Warehouse)
> mean.old
[1] 7.894706
> round(c(mean.fresh - mean.old - qnorm(0.975)*1.5*sqrt(2/17),
+ mean.fresh - mean.old + qnorm(0.975)*1.5*sqrt(2/17)), 2)
[1] -1.65 0.37
```

Thus, one is 95% confident that the difference in mean hardness for fresh and warehoused apples falls in the interval $[-1.65, 0.37]$ kg/meter². Since this interval contains zero, one can say that there is essentially no difference between the hardnesses for fresh and warehoused apples. ■

Note that no internal S functions such as `t.test` that assume unknown variances to construct the confidence interval reported in Example 8.10 were used.

8.2.4 Confidence Interval for the Difference in Population Means when Sampling from Independent Normal Distributions with Known but Unequal Variances

Consider two independent normal populations, $N(\mu_X, \sigma_X)$ and $N(\mu_Y, \sigma_Y)$, with known but unequal variances σ_X^2 and σ_Y^2 , respectively. If one takes random samples of size n_X and n_Y , respectively, the confidence interval for $\mu_X - \mu_Y$ can be constructed from knowledge of the sampling distribution of the statistic $\bar{X} - \bar{Y}$. Since

$$\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}\right),$$

the $(1 - \alpha) \cdot 100\%$ confidence interval for $(\mu_X - \mu_Y)$ is

$$\begin{aligned}
 & CI_{1-\alpha}(\mu_X - \mu_Y | \sigma_X \neq \sigma_Y \text{ but known}) = \\
 & \left[(\bar{x} - \bar{y}) - z_{1-\alpha/2} \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}, (\bar{x} - \bar{y}) + z_{1-\alpha/2} \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} \right]. \quad (8.12)
 \end{aligned}$$

Example 8.11 Suppose random samples of sizes $n_X = 50$ and $n_Y = 46$ are drawn from normal populations with standard deviations of 4.5 and 6, respectively, such that

$$\sum_{i=1}^{n_X} x_i = 420 \quad \text{and} \quad \sum_{i=1}^{n_Y} y_i = 405.$$

Construct a 97% confidence interval for $\mu_X - \mu_Y$.

Solution: Given that

$$\bar{x} = \frac{\sum_{i=1}^{50} x_i}{50} = \frac{420}{50} = 8.4 \quad \text{and} \quad \bar{y} = \frac{\sum_{i=1}^{46} y_i}{46} = \frac{405}{46} = 8.8,$$

the 97% confidence interval for $\mu_X - \mu_Y$ is constructed using (8.12) as

$$\begin{aligned}
 CI_{0.97}(\mu_X - \mu_Y) &= \left[(8.4 - 8.8) - 2.17 \sqrt{\frac{(4.5)^2}{50} + \frac{6^2}{46}}, (8.4 - 8.8) + 2.17 \sqrt{\frac{(4.5)^2}{50} + \frac{6^2}{46}} \right] \\
 &= [-2.76, 1.96].
 \end{aligned}$$

Note that since zero is contained in the interval, one concludes μ_X is not significantly different from μ_Y . To construct the confidence interval with S, key in

```

> round(qnorm(0.985), 2)
[1] 2.17
> round(c((8.4 - 8.8) - qnorm(0.985)*sqrt((4.5)^2/50 + (6)^2/46),
+ (8.4 - 8.8) + qnorm(0.985)*sqrt((4.5)^2/50 + (6)^2/46)), 2)
[1] -2.76 1.96

```

So, one is 97% confident that $\mu_X - \mu_Y$ lies in $[-2.76, 1.96]$. ■

Example 8.12 ▷ *Confidence Interval for $\mu_X - \mu_Y$: Calculus* ◁ Table 8.4 on the next page and data frame **Calculus** provide the mathematical assessment scores for 36 students enrolled in a biostatistics course according to whether or not the students had successfully completed a calculus course prior to enrolling in the biostatistics course. Construct a 95% confidence interval for the difference in the means of the mathematical assessment scores for students who had successfully completed a calculus course (X) and of those who had not (Y). Assume the distributions for X and Y have variances of 25 and 144, respectively. Determine if it is advantageous to take calculus prior to taking the biostatistics course.

Solution: Before using the confidence interval formula in (8.12), one needs to make sure the assumption of normality is satisfied. Consequently, a normal quantile-quantile plot for the values reported in Table 8.4 on the facing page was constructed and is shown in Figure 8.6 on the next page.

Table 8.4: Mathematical assessment scores for students enrolled in a biostatistics course (Calculus)

Y No Calculus						X Calculus					
73	39	55	72	88	64	82	90	85	87	86	79
57	58	75	44	76	68	85	92	89	82	92	82
64	55	62	61	76	40	85	87	92	85	95	90

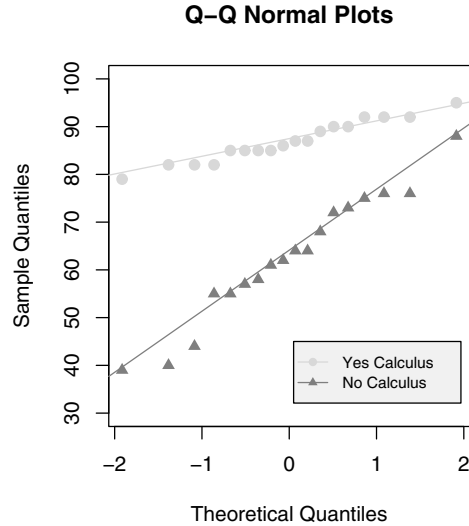


FIGURE 8.6: Superimposed normal quantile-quantile plots of the mathematical assessment scores for students enrolled in a biostatistics course who had successfully completed calculus and the mathematical assessment scores for students who had not successfully completed calculus

Since the points in Figure 8.6 fall relatively close to the straight lines, one decides the normality assumptions for using (8.12) are satisfied and continues by calculating the sample means for students who successfully completed calculus and those who have not yet successfully completed calculus as

$$\bar{x} = \frac{\sum_{i=1}^{18} x_i}{18} = \frac{1565}{18} = 86.94 \quad \text{and} \quad \bar{y} = \frac{\sum_{i=1}^{18} y_i}{18} = \frac{1127}{18} = 62.61.$$

The 95% confidence interval for $(\mu_X - \mu_Y)$ is constructed using (8.12) as

$$\begin{aligned} CI_{0.95}(\mu_X - \mu_Y) &= \\ &= \left[(86.94 - 62.61) - (1.96)\sqrt{\frac{25}{18} + \frac{144}{18}}, (86.94 - 62.61) + (1.96)\sqrt{\frac{25}{18} + \frac{144}{18}} \right] \\ &= [18.33, 30.34]. \end{aligned}$$

To construct the confidence interval with S, type

```
> attach(Calculus)
> str(Calculus)      # str ONLY works in R
'data.frame':  18 obs. of  2 variables:
 $ Yes.Calculus: num  82 90 85 87 86 79 85 92 89 82 ...
 $ No.Calculus : num  73 39 55 72 88 64 57 58 75 44 ...
> mean(Yes.Calculus)
[1] 86.94444
> mean(No.Calculus)
[1] 62.61111
> round(c(mean(Yes.Calculus) - mean(No.Calculus)
+       - qnorm(0.975)*sqrt(25/18+144/18),
+       mean(Yes.Calculus) - mean(No.Calculus)
+       + qnorm(0.975)*sqrt(25/18+144/18)),2)
[1] 18.33 30.34
```

Therefore, one is 95% confident that the difference in mean assessment scores for students who have successfully completed calculus prior to enrolling in biostatistics and those students who have not successfully completed calculus prior to enrolling in biostatistics lies in [18.33, 30.34]. It is advantageous to take calculus prior to taking biostatistics.

Note, once again, that the internal S function `t.test` was not used to construct the confidence interval since `t.test` assumes one is working with unknown variances; and in Example 8.12, the variances are known. If σ is unknown, use (8.16) on page 310. ■

8.2.5 Confidence Interval for the Difference in Means when Sampling from Independent Normal Distributions with Variances That Are Unknown but Assumed Equal

Suppose random samples of size n_X and n_Y , respectively, are taken from two normal distributions $N(\mu_X, \sigma)$ and $N(\mu_Y, \sigma)$, where σ is unknown. To obtain a confidence interval for $\mu_X - \mu_Y$, take advantage of Theorem 6.4 on page 237, which allows the use of the pivot

$$Q(\mathbf{X}, \mathbf{Y}; \mu_X - \mu_Y) = \frac{[(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)]}{\sqrt{S_p^2 \left(\frac{1}{n_X} + \frac{1}{n_Y} \right)}} \sim t_{n_X + n_Y - 2}. \quad (8.13)$$

The denominator of the pivotal quantity in (8.13) is an estimator for the variance of $\bar{X} - \bar{Y}$, where

$$S_p^2 = \frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2}. \quad (8.14)$$

Note that S_p^2 is a pooled estimator of the variance that weights the contributions of S_X^2 and S_Y^2 in proportion to the respective sample sizes n_X and n_Y . The degrees of freedom $n_X + n_Y - 2$ are denoted ν_p in the $(1 - \alpha) \cdot 100\%$ confidence interval for $\mu_X - \mu_Y$ given in (8.15).

$$\boxed{CI_{1-\alpha}(\mu_X - \mu_Y | \text{Assuming } \sigma_X = \sigma_Y \text{ but unknown}) = \left[(\bar{x} - \bar{y}) - t_{1-\alpha/2; \nu_p} s_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}, (\bar{x} - \bar{y}) + t_{1-\alpha/2; \nu_p} s_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}} \right]} \quad (8.15)$$

Example 8.13 A random sample from a $N(\mu_X, \sigma)$ population is taken where

$$n_X = 15, \quad \sum_{i=1}^{15} x_i = 53, \quad \text{and} \quad \sum_{i=1}^{15} x_i^2 = 222.$$

Another random sample is taken from a $N(\mu_Y, \sigma)$ population independent of the first sample such that

$$n_Y = 11, \quad \sum_{i=1}^{11} y_i = 77, \quad \text{and} \quad \sum_{i=1}^{11} y_i^2 = 560.$$

Obtain a 95% confidence interval for $\mu_X - \mu_Y$ by assuming the true but unknown variances are equal.

Solution: The sample means and sample variances are calculated as

$$\begin{aligned} \bar{x} &= \frac{\sum_{i=1}^{n_X} x_i}{n_X} = \frac{53}{15} = 3.53, \quad s_X^2 = \frac{\sum_{i=1}^{n_X} x_i^2 - n_X \bar{x}^2}{n_X - 1} = \frac{222 - (15)(3.53)^2}{15 - 1} = 2.51, \\ \bar{y} &= \frac{\sum_{i=1}^{n_Y} y_i}{n_Y} = \frac{77}{11} = 7, \quad \text{and} \quad s_Y^2 = \frac{\sum_{i=1}^{n_Y} y_i^2 - n_Y \bar{y}^2}{n_Y - 1} = \frac{560 - (11)(7)^2}{11 - 1} = 2.1. \end{aligned}$$

The pooled variance is given by

$$s_p^2 = \frac{(15 - 1)(2.51) + (11 - 1)(2.1)}{24} = 2.34,$$

where $s_p = 1.53$. Keeping in mind that $t_{0.975;24} = 2.06$, the 95% confidence interval for $\mu_X - \mu_Y$ is constructed using (8.15) as

$$\begin{aligned} CI_{0.95}(\mu_X - \mu_Y) &= \left[(3.53 - 7) - (2.06)(1.53) \sqrt{\frac{1}{15} + \frac{1}{11}}, \right. \\ &\quad \left. (3.53 - 7) + (2.06)(1.53) \sqrt{\frac{1}{15} + \frac{1}{11}} \right] = [-4.72, -2.22]. \end{aligned}$$

To construct this confidence interval with S, type

```
> round(qt(0.975,24), 2)
[1] 2.06
> sp <- round(sqrt((14*2.51+10*2.1)/24),2)
> sp
[1] 1.53
> round(c((3.53 - 7) - qt(0.975,24)*sp*sqrt(1/15 + 1/11),
+ (3.53 - 7) + qt(0.975,24)*sp*sqrt(1/15 + 1/11)),2)
[1] -4.72 -2.22
```

That is, one is 95% confident that the difference of means lies in $[-4.72, -2.22]$. ■

Example 8.14 Given the information from Example 8.10 on page 303, construct a 95% confidence interval for the difference in hardness between fresh and warehoused apples. Assume the samples come from normal and independent distributions with unknown but equal variances.

Solution: According to the solution for Example 8.10, the sample means for fresh and warehoused apples are $\bar{x} = 7.25$ and $\bar{y} = 7.89$, respectively. Next, calculate the respective sample variances as

$$s_X^2 = \frac{\sum_{i=1}^{n_X} (x_i - \bar{x})^2}{n_X - 1} = \frac{\sum_{i=1}^{17} (x_i - 7.25)^2}{16} = 1.51 \quad \text{and}$$

$$s_Y^2 = \frac{\sum_{i=1}^{n_Y} (y_i - \bar{y})^2}{n_Y - 1} = \frac{\sum_{i=1}^{17} (y_i - 7.89)^2}{16} = 1.79.$$

Note that the t -distribution has $n_X + n_Y - 2 = 17 + 17 - 2 = 32$ degrees of freedom and s_p is calculated as

$$s_p = \sqrt{\frac{16(1.51) + 16(1.79)}{32}} = 1.28.$$

Finally, the 95% confidence interval for $\mu_X - \mu_Y$ is calculated as

$$CI_{0.95}(\mu_X - \mu_Y) = \left[(7.25 - 7.89) - (2.04)(1.28)\sqrt{\frac{1}{17} + \frac{1}{17}}, \right. \\ \left. (7.25 - 7.89) + (2.04)(1.28)\sqrt{\frac{1}{17} + \frac{1}{17}} \right] = [-1.54, 0.26].$$

Assuming the data frame `Apple` is attached, this confidence interval can be constructed with S by keying in

```
> t.test(Fresh, Warehouse, var.equal=TRUE)$conf
[1] -1.5382253 0.2570488
attr(,"conf.level"):
[1] 0.95
```

So, one is 95% confident the difference in means for fresh and warehoused apple hardness falls in $[-1.54, 0.26]$ kg/meter². ■

8.2.6 Confidence Interval for a Difference in Means when Sampling from Independent Normal Distributions with Variances That Are Unknown and Unequal

If random samples of size n_X and n_Y are drawn from two independent normal distributions, say $N(\mu_X, \sigma_X)$ and $N(\mu_Y, \sigma_Y)$, where σ_X and σ_Y are unknown and unequal, a $(1 - \alpha) \cdot 100\%$ confidence interval for $\mu_X - \mu_Y$ is given by

$$CI_{1-\alpha}(\mu_X - \mu_Y | \text{Unknown } \sigma_X \neq \sigma_Y) = \left[(\bar{x} - \bar{y}) - t_{1-\alpha/2; \nu} \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}, (\bar{x} - \bar{y}) + t_{1-\alpha/2; \nu} \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}} \right] \quad (8.16)$$

The degrees of freedom, ν , for (8.16) are determined by (8.17). When ν does not give an integer value, it is truncated to give a conservative approximation. “Conservative” means

that the resulting confidence interval will have a confidence level of at least $1 - \alpha$.

$$\nu = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\frac{(s_X^2/n_X)^2}{n_X - 1} + \frac{(s_Y^2/n_Y)^2}{n_Y - 1}} \quad (8.17)$$

The standardized test statistic in (8.18) is used to construct a confidence interval for $\mu_X - \mu_Y$. The sampling distribution of (8.18) is very complicated, but Welch's approximation of a t_ν -distribution provides adequate results and will be used in this text:

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}} \sim t_\nu \quad (8.18)$$

Example 8.15 Suppose a random sample is taken from a $N(\mu_X, \sigma_X)$ population where

$$n_X = 15, \quad \sum_{i=1}^{15} x_i = 63, \quad \text{and} \quad \sum_{i=1}^{15} x_i^2 = 338.$$

A second random sample is taken from a $N(\mu_Y, \sigma_Y)$ population independent from the first sample such that

$$n_Y = 11, \quad \sum_{i=1}^{11} y_i = 66.4, \quad \text{and} \quad \sum_{i=1}^{11} y_i^2 = 486.$$

Construct a 95% confidence interval for $\mu_X - \mu_Y$ assuming the variances for the two populations are unknown and unequal.

Solution: Start by calculating the sample means and sample variances for the respective samples as well as ν , the value for the degrees of freedom:

$$\begin{aligned} \bar{x} &= \frac{63}{15} = 4.2 & s_X^2 &= \frac{\sum_{i=1}^{n_X} x_i^2 - n_X \bar{x}^2}{n_X - 1} = \frac{338 - (15)(4.2)^2}{15 - 1} = 5.24 \\ \bar{y} &= \frac{66.4}{11} = 6.04 & s_Y^2 &= \frac{\sum_{i=1}^{n_Y} y_i^2 - n_Y \bar{y}^2}{n_Y - 1} = \frac{486 - (11)(6.04)^2}{11 - 1} = 8.47 \end{aligned}$$

Next, (8.17) is used with the sample variances and respective sample sizes to determine ν :

$$\nu = \frac{\left(\frac{5.24}{15} + \frac{8.47}{11}\right)^2}{\frac{(5.24/15)^2}{14} + \frac{(8.47/11)^2}{10}} = 18.43 \approx 18$$

The 95% confidence interval for $\mu_X - \mu_Y$ is constructed using (8.16) as

$$\begin{aligned} CI_{0.95}(\mu_X - \mu_Y) &= \left[(4.2 - 6.04) - t_{0.975;18} \sqrt{\frac{5.24}{15} + \frac{8.47}{11}}, \right. \\ &\quad \left. (4.2 - 6.04) + t_{0.975;18} \sqrt{\frac{5.24}{15} + \frac{8.47}{11}} \right] \\ &= [-1.84 - (2.01)(1.06), -1.84 + (2.01)(1.06)] \\ &= [-4.06, 0.38]. \end{aligned}$$

To find this confidence interval with S, enter

```
> round(qt(0.975,18), 2)
[1] 2.01
> round(c((4.2 - 6.04) + qt(0.025,18)*sqrt(5.24/15 + 8.47/11),
+ (4.2 - 6.04) + qt(0.975,18)*sqrt(5.24/15 + 8.47/11)), 2)
[1] -4.06 0.38
```

One is 95% confident the difference of means lies in $[-4.06, 0.38]$. ■

Example 8.16 Using the information from Example 8.12, which provided the mathematical assessment scores for students enrolled in a biostatistics course according to whether they had completed a calculus course prior to enrolling in the biostatistics course, construct a 95% confidence interval for $\mu_X - \mu_Y$ assuming the samples are taken from distributions where the variances are unknown and unequal ($\sigma_X^2 \neq \sigma_Y^2$).

Solution: Recall from Example 8.12 that $\bar{x} = 86.94$ and $\bar{y} = 62.61$. Also recall that the assumption of normality for these data seemed plausible based on the normal quantile-quantile plot provided in Figure 8.6 on page 307. The respective sample variances are

$$s_X^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{18} (x_i - 86.94)^2}{17} = 18.64 \quad \text{and}$$

$$s_Y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^{18} (y_i - 62.61)^2}{17} = 174.84.$$

Next, (8.17) on the previous page is used with the sample variances and respective sample sizes to determine ν :

$$\nu = \frac{\left(\frac{18.64}{18} + \frac{174.84}{18}\right)^2}{\frac{(18.64/18)^2}{17} + \frac{(174.84/18)^2}{17}} = 20.58 \approx 20$$

The 95% confidence interval for $\mu_X - \mu_Y$ is constructed using (8.16) as

$$\begin{aligned} CI_{0.95}(\mu_X - \mu_Y) &= \left[(86.94 - 62.61) - t_{0.975;20} \sqrt{\frac{18.64}{18} + \frac{174.84}{18}}, \right. \\ &\quad \left. (86.94 - 62.61) + t_{0.975;20} \sqrt{\frac{18.64}{18} + \frac{174.84}{18}} \right] \\ &= [24.33 - (2.09)(3.28), 24.33 + (2.09)(3.28)] \\ &= [17.48, 31.19]. \end{aligned}$$

Assuming the data frame `Calculus` is attached, the confidence interval can be constructed directly with

```
> t.test(Yes.Calculus, No.Calculus, var.equal=FALSE)$conf
[1] 17.50677 31.15990
attr(,"conf.level")
[1] 0.95
```

One is 95% confident that the difference of means lies in $[17.48, 31.19]$. Note that S can compute quantiles in the t -distribution with non-integer degrees of freedom. In particular,

S uses the exact value for ν from (8.17) to find the critical value $t_{1-\alpha/2;\nu}$ in its confidence interval computation rather than truncating the value of ν . Consequently, the confidence interval computed with 20 degrees of freedom is slightly wider than the confidence interval S computes. ■

When working with normal distributions that have unknown variances, not pooling the variances and using (8.16) is generally the better method when the sample sizes are the same. It is also better when the sample sizes are unequal and the larger variance is associated with the larger sample size. Pooling the variances and using (8.15) should only be done if one is relatively confident that the variances are equal or if the larger variance is associated with the smaller sample size. For a summary of these methods, see Table 8.5.

Table 8.5: Methods for analyzing normal data

Condition	Method	Equation
Same Sample Sizes	Do Not Pool Variances	(8.16)
Larger Variance with Larger Sample	Do Not Pool Variances	(8.16)
Variances Equal	Pool Variances	(8.15)
Larger Variance with Smaller Sample	Pool Variances	(8.15)

8.2.7 Confidence Interval for the Mean Difference when the Differences Have a Normal Distribution

Information from two dependent distributions is often called **paired** or **dependent data**. Paired samples have some common intrinsic features such as members of the same family, animals from the same litter, etc. Data are also considered to be paired when the same sample is observed at different times. For example, suppose one is interested in evaluating the time undergraduate economic majors spend studying the first month of the semester and how much time they spend studying the last month of the semester. To help in the analysis, record the total time students spend studying the first and last months of the semester. This information is considered paired data since there are two measurements on each student. Scores recorded from a pre-test and post-test on the same group of people are also considered to be a paired or a dependent sample. In general, when the researcher is presented with paired samples, the standard approach is to analyze the differences between the paired data. In other words, if the population of pairs is $((X_1, Y_1), (X_2, Y_2), \dots)$, analyze the paired differences $D = (X_1 - Y_1, X_2 - Y_2, \dots)$. When there is a paired sample of size n_D , denote the sample differences as $d = (x_1 - y_1, \dots, x_{n_D} - y_{n_D})$. Provided the distribution of population differences is

$$D \sim N(\mu_D = \mu_X - \mu_Y, \sigma_D), \quad (8.19)$$

a confidence interval formula for μ_D when σ_D is unknown can be constructed using the pivotal quantity

$$Q(\mathbf{X}; \mu_D) = \frac{\bar{D} - \mu_D}{S_D / \sqrt{n_D}} \sim t_{n-1}, \quad (8.20)$$

where n_D represents the number of pairs in the sample and S_D is the standard deviation of the differences. Using (8.20) as a pivot, a $(1 - \alpha) \cdot 100\%$ confidence interval for the difference

in two dependent population means, $\mu_X - \mu_Y$, is given as

$$CI_{1-\alpha}(\mu_X - \mu_Y) = CI_{1-\alpha}(\mu_D) = \left[\bar{d} - t_{1-\alpha/2; n_D-1} \frac{s_D}{\sqrt{n_D}}, \bar{d} + t_{1-\alpha/2; n_D-1} \frac{s_D}{\sqrt{n_D}} \right] \quad (8.21)$$

Example 8.17 To compare the speed differences between two different brands of workstations (Sun and Digital), the times each brand took to complete complex simulations were recorded. Five complex simulations were selected, and the five selected simulations were run on both workstations. The resulting times in minutes for the five simulations are given in Table 8.6 and stored in data frame `Sundig`. Construct a 95% confidence interval for μ_D , the average time difference between SUN and DIGITAL workstations. Is one of the workstations faster than the other?

Table 8.6: Time to complete a complex simulation in minutes (`Sundig`)

Simulation	SUN	DIGITAL	Difference
1	110	102	8
2	125	120	5
3	141	135	6
4	113	114	-1
5	182	175	7

Solution: Since each one of the five selected complex simulations was run on both workstations, these samples are dependent. The differences between the dependent samples are $d = (8, 5, 6, -1, 7)$, $\bar{d} = 5$ minutes, and $s_D = 3.53$ minutes. Before using (8.21), one needs to verify the distribution of differences is normal. To check the normality assumption, use the functions `qqnorm()` and `qqline()` on the sample differences, d . The resulting normal quantile-quantile plot is shown in Figure 8.7 on the next page. Based on Figure 8.7, it is not immediately clear that the distribution of differences is normal due to the outlier in the lower left of the plot. At this point, one should look at several normal quantile-quantile plots for samples of size five using the `ntester()` function. The results of using the function `ntester()` on the sample differences are shown in Figure 8.8 on the facing page. After using the `ntester()` function on the differences and viewing the output in Figure 8.8, one can conclude that it is not unreasonable to assume the distribution of differences between Sun and Digital workstations follow a normal distribution and can use (8.21) to construct the 95% confidence interval for $\mu_D = \mu_{\text{SUN}} - \mu_{\text{DIG}}$ as follows:

$$\begin{aligned} CI_{0.95}(\mu_{\text{SUN}} - \mu_{\text{DIG}}) &= \left[\bar{d} - t_{0.975; n_D-1} \frac{s_D}{\sqrt{n_D}}, \bar{d} + t_{0.975; n_D-1} \frac{s_D}{\sqrt{n_D}} \right] \\ &= \left[5 - (2.78) \frac{3.53}{\sqrt{5}}, 5 + (2.78) \frac{3.53}{\sqrt{5}} \right] = [0.61, 9.39]. \end{aligned}$$

One is 95% confident μ_D lies in [0.61, 9.39] minutes. Since the confidence interval does not contain zero, one can be 95% confident that $\mu_D = \mu_{\text{SUN}} - \mu_{\text{DIG}} > 0$. This implies that $\mu_{\text{SUN}} > \mu_{\text{DIG}}$, which means that the Digital workstation is faster than the Sun workstation.

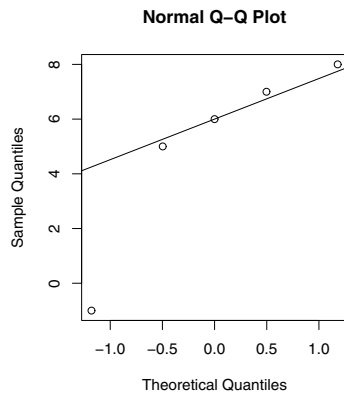


FIGURE 8.7: Normal quantile-quantile plot of the time differences between Sun and Digital workstations to complete complex simulations

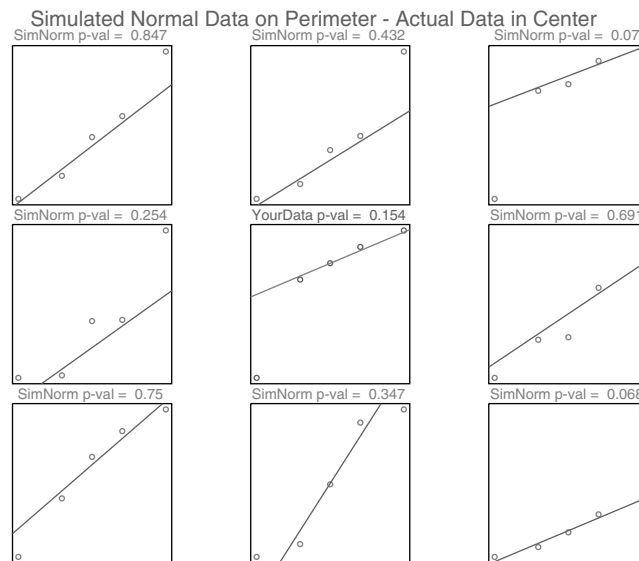


FIGURE 8.8: Quantile-quantile plot of the time differences between Sun and Digital workstations to complete complex simulations shown in the middle with normal quantile-quantile plots of random normal data depicted on the outside plots

To verify the value $t_{0.975;4}$ and to calculate a 95% confidence interval for μ_D with S , enter

```
> round(qt(0.025,4), 2)
[1] -2.78
> attach(Sundig)
> t.test(SUN, DIGITAL, paired=TRUE)$conf
[1] 0.6100548 9.3899452
attr("conf.level")
[1] 0.95
```



8.3 Confidence Intervals for Population Variances

8.3.1 Confidence Interval for the Population Variance of a Normal Population

This section considers a normal population $N(\mu, \sigma)$ from which a random sample of size n is taken. The confidence interval for σ^2 is based on the pivot

$$Q(\mathbf{X}; \sigma^2) = \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2 \quad (8.22)$$

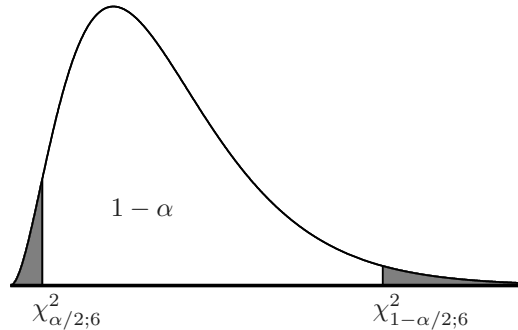


FIGURE 8.9: Chi-square distribution with six degrees of freedom depicting the points $\chi_{\alpha/2;6}^2$ and $\chi_{1-\alpha/2;6}^2$

The pivotal quantity (8.22) is not very robust with respect to the normality assumption. Consequently, before constructing a confidence interval for σ^2 , one should always check the sample for normality using a graphical procedure such as a normal quantile-quantile plot (`qqnorm()`). Although Pearson's χ^2 distribution is not symmetric (see Figure 8.9), one can use the sampling distribution of the statistic $(n-1)S^2/\sigma^2$ and the definition of percentiles to obtain

$$\mathbb{P}\left(\chi_{\alpha/2;n-1}^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{1-\alpha/2;n-1}^2\right) = 1 - \alpha. \quad (8.23)$$

To arrive at the standard confidence interval form for the variance, first take the reciprocal inside the probability statement of (8.23) as shown in (8.24). Then, multiply everything inside the probability statement of (8.24) by $(n-1)S^2$ to obtain the probability statement shown in (8.25):

$$\mathbb{P}\left(\frac{1}{\chi_{\alpha/2;n-1}^2} \geq \frac{\sigma^2}{(n-1)S^2} \geq \frac{1}{\chi_{1-\alpha/2;n-1}^2}\right) = 1 - \alpha, \quad (8.24)$$

$$\mathbb{P}\left(\frac{(n-1)S^2}{\chi_{1-\alpha/2;n-1}^2} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{\alpha/2;n-1}^2}\right) = 1 - \alpha. \quad (8.25)$$

Using the probability statement (8.25) at a fixed confidence level of $(1 - \alpha)$, the standard form for the confidence interval for σ^2 is illustrated in (8.26). Note that the confidence interval for the variance is not centered around the point estimate (the sample variance, s^2).

$$CI_{1-\alpha}(\sigma^2) = \left[\frac{(n-1)s^2}{\chi_{1-\alpha/2; n-1}^2}, \frac{(n-1)s^2}{\chi_{\alpha/2; n-1}^2} \right]. \quad (8.26)$$

Example 8.18 Construct an 80% confidence interval for σ_X^2 using the information from Example 8.15.

Solution: Recall that the underlying distribution in Example 8.15 was assumed to be $N(\mu_X, \sigma_X)$ and the sample information provided revealed that $n_X = 15$, $\bar{x} = 4.2$, and $s_X^2 = 5.24$. Using (8.26), the 80% confidence interval for σ_X^2 is calculated as

$$\begin{aligned} CI_{0.8}(\sigma_X^2) &= \left[\frac{(n_X - 1)s_X^2}{\chi_{0.9; n-1}^2}, \frac{(n_X - 1)s_X^2}{\chi_{0.1; n-1}^2} \right] = \left[\frac{14(5.24)}{\chi_{0.9; 14}^2}, \frac{14(5.24)}{\chi_{0.1; 14}^2} \right] \\ &= \left[\frac{73.36}{21.06}, \frac{73.36}{7.79} \right] = [3.48, 9.42]. \end{aligned}$$

To construct this confidence interval with S, type

```
> round(qchisq(0.1,14), 2)
[1] 7.79
> qchisq(0.9,14)
[1] 21.06
> round(c(14*5.24/qchisq(0.9,14), 14*5.24/qchisq(0.1,14)), 2)
[1] 3.48 9.42
```

Therefore, one is 80% confident the variance falls in [3.48, 9.42]. ■

Example 8.19 The data frame `barley` is in the `lattice` package and contains yield, variety, year, and site, giving barley yields (bushels/acre) in 1931 and 1932 for 10 varieties of barley grown at six sites. The S-PLUS data frame `barley` is identical.

- Construct a 95% confidence interval for μ , the mean barley yield in 1932.
- Construct a 95% confidence interval for σ^2 , the variance of barley yield in 1932.

Solution: Start by looking at the distribution of 1932 barley yield using the functions `qqnorm()` and `qqline()` to create the normal quantile-quantile plot shown in Figure 8.10 on the following page. Since the values in Figure 8.10 on the next page are fairly linear, it is decided the assumptions to use both (8.10) and (8.26) are satisfied.

- To construct a 95% confidence interval for μ , use (8.10) as follows:

$$\begin{aligned} CI_{1-0.05}(\mu) &= \left[\bar{x} - t_{1-0.05/2; n-1} \frac{s}{\sqrt{n}}, \bar{x} + t_{1-0.05/2; n-1} \frac{s}{\sqrt{n}} \right] \\ &= \left[31.76 - (2.00) \frac{9.38}{\sqrt{60}}, 31.76 + (2.00) \frac{9.38}{\sqrt{60}} \right] \\ &= [29.34, 34.19] \end{aligned} \quad (8.27)$$

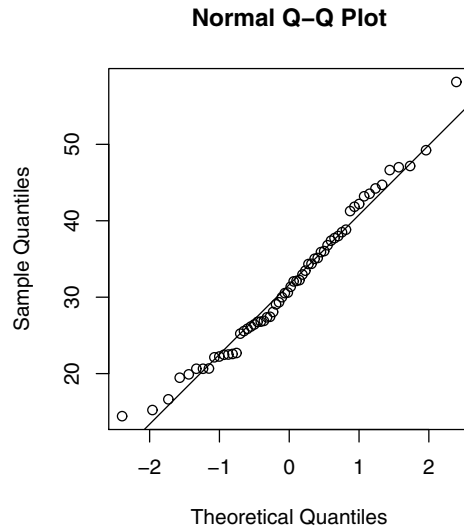


FIGURE 8.10: Quantile-quantile plot of 1932 barley yield in bushels/acre

```
> library(lattice) # Not needed for S-PLUS
> attach(barley)
> n <- length(yield[year==1932])
> n
[1] 60
> mean(yield[year==1932])
[1] 31.76333
> var(yield[year==1932])
[1] 88.06803
> sd(yield[year==1932]) #S-PLUS: stdev(yield[year==1932])
[1] 9.384457
> qt(.975, n-1)
[1] 2.000995
```

To construct the confidence interval directly with S, key in

```
> t.test(yield[year==1932])$conf
[1] 29.33907 34.18759
attr(,"conf.level"):
[1] 0.95
```

So, one is 95% confident that the mean barley yield (bushels/acre) lies in [29.34, 34.19].

(b) To construct a 95% confidence interval for σ^2 , use (8.26):

$$CI_{0.95}(\sigma^2) = \left[\frac{(n-1)s^2}{\chi_{0.975; n-1}^2}, \frac{(n-1)s^2}{\chi_{0.025; n-1}^2} \right] = \left[\frac{59(88.07)}{82.12}, \frac{59(88.07)}{39.66} \right] = [63.28, 131.01].$$

To verify the previous values and construct this confidence interval with S, enter

```
> s2<- var(yield[year==1932])
> s2
[1] 88.06803
> ChiL <- qchisq(.025,59)
> ChiL
[1] 39.66186
> ChiU <- qchisq(.975,59)
> ChiU
[1] 82.1174
> n <- length(yield[year==1932])
> n
[1] 60
> round(c((n-1)*s2/ChiU, (n-1)*s2/ChiL),2)
[1] 63.28 131.01
```

One is 95% confident that the variance of barley yield lies in [63.28, 131.01] (bushels/acre)². ■

8.3.2 Confidence Interval for the Ratio of Population Variances when Sampling from Independent Normal Distributions

Now consider the construction of confidence intervals for σ_X^2/σ_Y^2 when there are two normal and independent populations, $N(\mu_X, \sigma_X)$ and $N(\mu_Y, \sigma_Y)$, from which one takes random samples of sizes n_X and n_Y , respectively. The goal is to construct a confidence interval for the ratio of the variances, σ_X^2/σ_Y^2 . Generally, one is looking for 1 to be in the interval, indicating that the variances are equal. To construct a confidence interval for σ_X^2/σ_Y^2 , use Theorem 6.5 on page 239, which states that if one has two random samples X_1, \dots, X_{n_X} and Y_1, \dots, Y_{n_Y} that are taken from independent normal populations where $X \sim N(\mu_X, \sigma_X)$ and $Y \sim N(\mu_Y, \sigma_Y)$, then the random variable

$$\frac{S_Y^2/\sigma_Y^2}{S_X^2/\sigma_X^2} \sim F_{n_Y-1, n_X-1}. \quad (8.28)$$

By using (8.28), construct the $(1-\alpha)$ probability statement shown in (8.29) and graphically illustrated in Figure 8.11 on the next page for an F distribution with 10 and 10 degrees of freedom:

$$\mathbb{P}\left(f_{\alpha/2; n_Y-1, n_X-1} \leq \frac{S_Y^2/\sigma_Y^2}{S_X^2/\sigma_X^2} \leq f_{1-\alpha/2; n_Y-1, n_X-1}\right) = 1 - \alpha \quad (8.29)$$

After multiplying everything inside the probability statement given in (8.29) by $\frac{S_X^2}{S_Y^2}$, (8.30) is used to derive the final confidence interval statement given in (8.31):

$$\mathbb{P}\left(f_{\alpha/2; n_Y-1, n_X-1} \frac{S_X^2}{S_Y^2} \leq \frac{\sigma_X^2}{\sigma_Y^2} \leq f_{1-\alpha/2; n_Y-1, n_X-1} \frac{S_X^2}{S_Y^2}\right) = 1 - \alpha \quad (8.30)$$

$$\boxed{CI_{1-\alpha}\left(\frac{\sigma_X^2}{\sigma_Y^2}\right) = \left[f_{\alpha/2; n_Y-1, n_X-1} \frac{s_X^2}{s_Y^2}, f_{1-\alpha/2; n_Y-1, n_X-1} \frac{s_X^2}{s_Y^2}\right]} \quad (8.31)$$

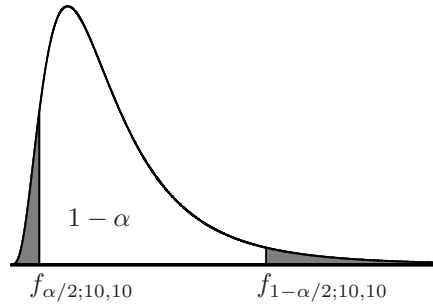


FIGURE 8.11: F distribution with ten and ten degrees of freedom depicting the points $f_{\alpha/2;10,10}$ and $f_{1-\alpha/2;10,10}$

For sheer convenience, denote the larger sample variance as s_X^2 when constructing a confidence interval for the ratio of two population variances. Consequently, the numerator for the ratio of the sample variances will always contain the larger of the two sample variances. Many tables involving the F distribution only provide values for percentiles in the right tail. However, this does not present a problem provided one remembers that values in the left tail of the F distribution can be found from the values in the right tail of an F distribution by using (8.32). Note that the order of the degrees of freedom changes in the reciprocal.

$$f_{\alpha/2;n_Y-1,n_X-1} = \frac{1}{f_{1-\alpha/2;n_X-1,n_Y-1}}. \quad (8.32)$$

Example 8.20 Using the information from Example 8.13 on page 309, construct a 90% confidence interval for the ratio of variances.

Solution: In Example 8.13, the larger sample variance, s_X^2 , was 2.51, $n_X = 15$, and the smaller sample variance, s_Y^2 , was 2.1, $n_Y = 11$. Consequently, the 90% confidence interval for the ratio of variances is constructed using (8.31) as shown in the following where $f_{0.05;10,14} = 0.35$ and $f_{0.95;10,14} = 2.60$:

$$CI_{0.9} \left(\frac{\sigma_X^2}{\sigma_Y^2} \right) = \left[f_{0.05;10,14} \frac{s_X^2}{s_Y^2}, f_{0.95;10,14} \frac{s_X^2}{s_Y^2} \right] = \left[(0.35) \frac{2.51}{2.1}, (2.60) \frac{2.51}{2.1} \right] = [0.42, 3.11] \quad (8.33)$$

To find $f_{0.05;10,14}$, $f_{0.95;10,14}$ and a 95% confidence interval for $\frac{\sigma_X^2}{\sigma_Y^2}$ with S, type

```
> round(qf(0.05,10,14), 2)
[1] 0.35
> round(qf(0.95,10,14), 2)
[1] 2.60
> round(c(qf(0.05,10,14)*(2.51/2.1), qf(0.95,10,14)*(2.51/2.1)), 2)
[1] 0.42 3.11
```

So, one is 90% confident the ratio of the variance lies in $[0.42, 3.11]$. Note that this interval includes 1, which indicates there is not evidence to suggest the variances are different. ■

Example 8.21 Given the information in Table 8.3 on page 304, construct a 95% confidence interval for the ratio of the variances.

Solution: According to Example 8.14, $s_X^2 = 1.51$ and $s_Y^2 = 1.79$. Also recall that in the solution to Example 8.10, a normal quantile-quantile plot was created and illustrated in Figure 8.5 on page 304 that justified the assumptions that both fresh and warehoused apples follow a normal distribution. Consequently, the appropriate confidence interval formula for the ratio of the variances is given in (8.31). However, since $s_Y^2 = 1.79$ and $s_X^2 = 1.51$, reverse s_X^2 for s_Y^2 in the confidence interval formula provided in (8.31) to construct a 95% confidence interval for the ratio of population variances:

$$\begin{aligned} CI_{0.95} \left(\frac{\sigma_Y^2}{\sigma_X^2} \right) &= \left[f_{0.025;16,16} \frac{s_Y^2}{s_X^2}, f_{0.975;16,16} \frac{s_Y^2}{s_X^2} \right] \\ &= [(0.36)(1.19), (2.76)(1.19)] = [0.43, 3.27]. \end{aligned} \quad (8.34)$$

To verify the previous values and to construct a 95% confidence interval for the ratio of variances with S, attach **Apple** and key in

```
> var(Warehouse)
[1] 1.790951
> var(Fresh)
[1] 1.510438
> round(var(Warehouse)/var(Fresh), 2)
[1] 1.19
> round(qf(0.025, 16, 16), 2)
[1] 0.36
> round(qf(0.975, 16, 16), 2)
[1] 2.76
> var.test(Warehouse, Fresh)$conf
[1] 0.429396 3.274189
attr(,"conf.level"):
[1] 0.95
```

One is 95% confident that the ratio of variances falls in [0.43, 3.27], which indicates that a pooled variance could be justified for confidence interval calculations regarding the means. ■

8.4 Confidence Intervals Based on Large Samples

Provided the sample size, n , is sufficiently large, one can take advantage of the asymptotic properties of maximum likelihood estimators to construct confidence intervals since, as $n \rightarrow \infty$,

$$\hat{\theta}(\mathbf{X}) \sim N \left(\theta, \sqrt{I_n(\theta)^{-1}} \right). \quad (8.35)$$

Using (8.35), one can construct asymptotic confidence intervals of the type given in (8.36). Note that $\sigma_{\hat{\theta}(\mathbf{X})}$ is the standard deviation of the estimator $\hat{\theta}(\mathbf{X})$. Specifically, in the multi-parameter case, $\sigma_{\hat{\theta}(\mathbf{X})}$ is the square root of the corresponding diagonal element of the inverse

of the information matrix. When $\sigma_{\hat{\theta}(\mathbf{X})}$ is unknown, the estimate $\hat{\sigma}_{\hat{\theta}(\mathbf{X})}$ is used in place of $\sigma_{\hat{\theta}(\mathbf{X})}$. Be sure to see that $\hat{\sigma}_{\hat{\theta}(\mathbf{X})}$ is calculated from the data \mathbf{x} .

$$CI_{1-\alpha}(\theta) = \left[\hat{\theta}(\mathbf{x}) - z_{1-\alpha/2} \cdot \sigma_{\hat{\theta}(\mathbf{X})}, \hat{\theta}(\mathbf{x}) + z_{1-\alpha/2} \cdot \sigma_{\hat{\theta}(\mathbf{X})} \right] \quad (8.36)$$

Example 8.22 Given a random sample of size 200 from an exponential distribution, find a 90% confidence interval for θ if it is true that

$$\sum_{i=1}^{200} x_i = 400.$$

As a reminder, the exponential distribution is

$$f(x, \theta) = \frac{1}{\theta} e^{-\frac{1}{\theta}x}, \quad x \geq 0, \quad \theta > 0 \quad (8.37)$$

Solution: The reader should verify that the maximum likelihood estimator of θ is $\hat{\theta}(\mathbf{X}) = \bar{X}$ and the variance of \bar{X} is $\frac{\theta^2}{n}$. (Hint: See Example 7.6 on page 250.) Because \bar{X} is the maximum likelihood estimator of θ , it follows that the maximum likelihood estimator of $\frac{\theta^2}{n}$ is $\frac{\bar{X}^2}{n}$ due to the invariance property of MLEs (property 2 on page 273). From the sample information, calculate

$$\hat{\theta}(\mathbf{x}) = \bar{x} = 2 \quad \text{and} \quad \hat{\sigma}_{\hat{\theta}(\mathbf{x})}^2 = \frac{\bar{x}^2}{n} = 0.02.$$

Given that the confidence level is 0.9, $z_{1-\alpha/2} = z_{0.95} = 1.64$, the 90% confidence interval for θ is constructed using (8.36):

$$CI_{0.90}(\theta) = \left[2 - 1.64\sqrt{0.02}, 2 + 1.64\sqrt{0.02} \right] = [1.77, 2.23]. \quad (8.38)$$

So, one is 90% confident the exponential parameter θ falls in $[1.77, 2.23]$. ■

8.4.1 Confidence Interval for the Population Proportion

The maximum likelihood estimator of the population proportion π is P , the sample proportion. See Example 7.15 on page 258 for the derivation of the maximum likelihood estimator of π . To calculate the Fisher information $I_n(\pi)$, use (7.49) on page 270. Since

$$\frac{\partial^2 \ln L(\pi|\mathbf{X})}{\partial \pi^2} = \frac{-\sum_{i=1}^n x_i}{\pi^2} - \frac{n - \sum_{i=1}^n x_i}{(1-\pi)^2} \quad (8.39)$$

from Example 7.15 on page 258, by multiplying (8.39) by -1 and taking the expected value of the result, gives

$$\begin{aligned} -E \left[\frac{\partial^2 \ln L(\pi|\mathbf{X})}{\partial \pi^2} \right] &= E \left[\frac{\sum_{i=1}^n x_i}{\pi^2} \right] + E \left[\frac{n - \sum_{i=1}^n x_i}{(1-\pi)^2} \right] \\ &= \frac{n\pi}{\pi^2} + \frac{n - n\pi}{(1-\pi)^2} \\ &= \frac{n}{\pi} + \frac{n}{1-\pi} = \frac{n}{\pi(1-\pi)} \end{aligned} \quad (8.40)$$

Consequently, the Fisher information, $I_n(\pi)^{-1}$, is given in (8.41).

$$I_n(\pi)^{-1} = \frac{\pi(1-\pi)}{n} \quad (8.41)$$

Taking advantage of the asymptotic properties of MLE estimators allows one to write

$$\hat{\pi}(\mathbf{X}) = P \sim N\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right) \text{ as } n \rightarrow \infty;$$

and using (8.36), one can construct a $(1-\alpha) \cdot 100\%$ asymptotic confidence interval for π as shown in (8.42) where $\hat{\sigma}_{\hat{\pi}(\mathbf{x})} = \sqrt{\frac{p(1-p)}{n}}$. The confidence interval in (8.42) can also be derived using the approximate sampling distribution of P from Section 6.5.3:

$$CI_{1-\alpha}(\pi) = \left[p - z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}}, p + z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}} \right] \quad (8.42)$$

A more accurate confidence interval for π can be obtained by solving for the values that satisfy (8.43) instead of replacing $\sigma_{\hat{\pi}(\mathbf{x})}$ with its MLE $\hat{\sigma}_{\hat{\pi}(\mathbf{x})}$. Solving for the values that satisfy (8.43) is slightly more involved but produces the confidence interval given in (8.44). Recent research (Agresti and Coull, 1998) shows that the confidence interval in (8.44) can be used for a wide range of parameters and sample sizes. Therefore, when working with smaller sample sizes, the confidence interval formula in (8.44) is preferred over the confidence interval formula (8.42) as it returns confidence intervals whose nominal confidence level is closer to the user specified $1-\alpha$ level. If the sample size is large, $z_{1-\alpha/2}^2/2n$ is negligible compared to p , $z_{1-\alpha/2}^2/4n^2$ under the square root is negligible compared to $p(1-p)/n$, and $z_{1-\alpha/2}^2/n$ is negligible compared to 1. If the negligible terms are ignored, the confidence interval formula in (8.42) emerges.

$$\mathbb{P}\left(P - z_{1-\alpha/2} \sqrt{\frac{\pi(1-\pi)}{n}} \leq \pi \leq P + z_{1+\alpha/2} \sqrt{\frac{\pi(1-\pi)}{n}}\right) = 1 - \alpha \quad (8.43)$$

$$CI_{1-\alpha}(\pi) = \left[\frac{p + \frac{z_{1-\alpha/2}^2}{2n} - z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n} + \frac{z_{1-\alpha/2}^2}{4n^2}}}{\left(1 + \frac{z_{1-\alpha/2}^2}{n}\right)}, \frac{p + \frac{z_{1-\alpha/2}^2}{2n} + z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n} + \frac{z_{1-\alpha/2}^2}{4n^2}}}{\left(1 + \frac{z_{1-\alpha/2}^2}{n}\right)} \right] \quad (8.44)$$

When S uses (8.44) to construct confidence intervals, under certain conditions, it also applies a Yates' continuity correction to p so that the p used in the lower limit is $p_L = p - \frac{1}{2n}$ and the p used in the upper limit is $p_U = p + \frac{1}{2n}$.

Example 8.23 A professor is interested in what percent of students pass an introductory statistics class. He takes a random sample of 40 introductory statistics students and finds that 26 passed. Help the professor construct 95% confidence intervals for the true percent of students who pass using

- The asymptotic confidence interval for π based on the MLE of $\hat{\sigma}_{\hat{\pi}(x)}$ given in (8.42).
- The preferred confidence interval for smaller sample sizes given in (8.44).
- The preferred confidence interval for smaller sample sizes with continuity corrections applied to the ps (use p_L and p_U).

Solution: Because all of the confidence intervals are to have 95% confidence, $z_{1-\alpha/2} = z_{1-0.05/2} = z_{0.975} = 1.96$. The sample proportion is $p = \frac{26}{40} = 0.65$.

- The asymptotic confidence interval for π is

$$\begin{aligned} CI_{0.95}(\pi) &= \left[p - z_{0.975} \sqrt{\frac{p(1-p)}{n}}, p + z_{0.975} \sqrt{\frac{p(1-p)}{n}} \right] \\ &= \left[0.65 - 1.96 \sqrt{\frac{(0.65)(1-0.65)}{40}}, 0.65 + 1.96 \sqrt{\frac{(0.65)(1-0.65)}{40}} \right] \\ &= [0.502, 0.798] \end{aligned}$$

To compute this interval with S, enter

```
> p <- 26/40
> z <- qnorm(.975)
> n <- 40
> round(z,2)           # z_(0.975)
[1] 1.96
> round(c(p - z*sqrt(p*(1 - p)/n), p + z*sqrt(p*(1 - p)/n)),3)
[1] 0.502 0.798
```

- The preferred confidence interval for smaller sample sizes is

$$\begin{aligned} CI_{0.95}(\pi) &= \left[\frac{p + \frac{z_{1-\alpha/2}^2}{2n} - z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n} + \frac{z_{1-\alpha/2}^2}{4n^2}}}{\left(1 + \frac{z_{1-\alpha/2}^2}{n}\right)}, \right. \\ &\quad \left. \frac{p + \frac{z_{1-\alpha/2}^2}{2n} + z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n} + \frac{z_{1-\alpha/2}^2}{4n^2}}}{\left(1 + \frac{z_{1-\alpha/2}^2}{n}\right)} \right] \\ &= \left[\frac{0.65 + \frac{1.96^2}{(2)(40)} - 1.96 \sqrt{\frac{0.65(1-0.65)}{40} + \frac{1.96^2}{(4)(40^2)}}}{\left(1 + \frac{1.96^2}{40}\right)}, \right. \\ &\quad \left. \frac{0.65 + \frac{1.96^2}{(2)(40)} + 1.96 \sqrt{\frac{0.65(1-0.65)}{40} + \frac{1.96^2}{(4)(40^2)}}}{\left(1 + \frac{1.96^2}{40}\right)} \right] = [0.495, 0.779] \end{aligned}$$

To compute this interval with S, key in

```
> round(prop.test(26,40, correct=FALSE)$conf,3)
[1] 0.495 0.779
attr(,"conf.level")
[1] 0.95
```

(c) Using the values $p_L = p - \frac{1}{2n} = 0.65 - \frac{1}{(2)(40)} = 0.6375$ and $p_U = p + \frac{1}{2n} = 0.65 + \frac{1}{(2)(40)} = 0.6625$, the confidence interval is

$$\begin{aligned}
 CI_{0.95}(\pi) &= \left[\frac{p_L + \frac{z_{1-\alpha/2}^2}{2n} - z_{1-\alpha/2} \sqrt{\frac{p_L(1-p_L)}{n} + \frac{z_{1-\alpha/2}^2}{4n^2}}}{\left(1 + \frac{z_{1-\alpha/2}^2}{n}\right)}, \right. \\
 &\quad \left. \frac{p_U + \frac{z_{1-\alpha/2}^2}{2n} + z_{1-\alpha/2} \sqrt{\frac{p_U(1-p_U)}{n} + \frac{z_{1-\alpha/2}^2}{4n^2}}}{\left(1 + \frac{z_{1-\alpha/2}^2}{n}\right)} \right] \\
 &= \left[\frac{0.6375 + \frac{1.96^2}{(2)(40)} - 1.96 \sqrt{\frac{0.6375(1-0.6375)}{40} + \frac{1.96^2}{(4)(40^2)}}}{\left(1 + \frac{1.96^2}{40}\right)}, \right. \\
 &\quad \left. \frac{0.6625 + \frac{1.96^2}{(2)(40)} + 1.96 \sqrt{\frac{0.6625(1-0.6625)}{40} + \frac{1.96^2}{(4)(40^2)}}}{\left(1 + \frac{1.96^2}{40}\right)} \right] = [0.483, 0.789]
 \end{aligned}$$

The interval with continuity correction is computed with S by typing

```
> round(prop.test(26,40, correct=TRUE)$conf,3)
[1] 0.483 0.789
attr(,"conf.level")
[1] 0.95
```

So, depending on which confidence interval the professor prefers, he can be 95% confident that the proportion of students who pass lies in $[0.502, 0.798]$, $[0.495, 0.779]$, or $[0.483, 0.789]$. ■

Example 8.24 A computer firm would like to construct three confidence intervals for the proportion of supermarkets that use a computerized database to manage their warehouses. Suppose 200 supermarkets are surveyed and 157 of the 200 supermarkets have computerized inventories. Construct 90%, 95%, and 99% confidence intervals for the true proportion of supermarkets that use a computerized database to manage the inventory of their warehouses.

Solution: Since $p = \frac{157}{200} = 0.785$, a $(1 - \alpha) \cdot 100\%$ confidence interval for π can be constructed using (8.42) as follows:

$$CI_{1-\alpha}(\pi) = \left[0.785 - z_{1-\alpha/2} \sqrt{\frac{(0.785)(0.215)}{200}}, 0.785 + z_{1-\alpha/2} \sqrt{\frac{(0.785)(0.215)}{200}} \right]. \quad (8.45)$$

Making appropriate substitutions for $z_{1-\alpha/2}$ in (8.45) yields

$$\begin{aligned} CI_{0.90}(\pi) &= [0.785 - 1.645(0.03), 0.785 + 1.645(0.03)] = [0.74, 0.83] \\ CI_{0.95}(\pi) &= [0.785 - 1.960(0.03), 0.785 + 1.960(0.03)] = [0.73, 0.84] \\ CI_{0.99}(\pi) &= [0.785 - 2.576(0.03), 0.785 + 2.576(0.03)] = [0.71, 0.86] \end{aligned}$$

The computer firm is 90% confident the population proportion of supermarkets that use a computerized database to manage their warehouses lies in $[0.74, 0.83]$, 95% confident this population proportion lies in $[0.73, 0.84]$, and 99% confident this population proportion lies in $[0.71, 0.86]$. Take special note that the widths of the confidence intervals increase as the confidence level increases. To find these confidence intervals using the confidence interval formula in (8.44) with S, enter

```
> round(prop.test(157, 200, conf.level=0.90, correct=FALSE)$conf, 2)
[1] 0.73 0.83
> round(prop.test(157, 200, conf.level=0.95, correct=FALSE)$conf, 2)
[1] 0.72 0.84
> round(prop.test(157, 200, conf.level=0.99, correct=FALSE)$conf, 2)
[1] 0.70 0.85
```

Example 8.25 ▷ *Confidence Interval and Sample Size for π* ◁ The Department of Agriculture wants to estimate the proportion of rural farm owners that are under 40 years of age. They take a random sample of 2000 farms and find that 400 of the 2000 owners are under the age of 40.

- Construct a 95% confidence interval for π using the asymptotic confidence interval for π based on the MLE of $\hat{\sigma}_{\hat{\pi}(\mathbf{x})}$ given in (8.42).
- Determine the required sample size so that the maximum margin of error is within 0.015 of the true value of π for a 95% confidence level.

Solution: Note that $p = \frac{400}{2000} = 0.20$.

- A 95% confidence interval for π using (8.42) is

$$\begin{aligned} CI_{0.95}(\pi) &= \left[p - z_{1-0.05/2} \sqrt{\frac{p(1-p)}{n}}, p + z_{1-0.05/2} \sqrt{\frac{p(1-p)}{n}} \right] \\ &= \left[0.2 - (1.96) \sqrt{\frac{0.2(1-0.2)}{2000}}, 0.2 + (1.96) \sqrt{\frac{0.2(1-0.2)}{2000}} \right] \\ &= [0.182, 0.218]. \end{aligned}$$

To verify the confidence interval with S, type

```
> p <- 400/2000
> n <- 2000
> z <- qnorm(.975)
> round(z, 2)
[1] 1.96
> round(c(p-z*sqrt(p*(1-p)/n), p+z*sqrt(p*(1-p)/n)), 3)
[1] 0.182 0.218
```

(b) In order to construct a confidence interval such that the maximum margin of error does not exceed 0.015, one needs to ensure that

$$(1.96)\sqrt{\frac{p(1-p)}{n}} < 0.015. \quad (8.46)$$

To maximize the margin of error, use $p = \frac{1}{2}$ regardless of any prior information concerning p . Using a value for p of $\frac{1}{2}$ will ensure the margin of error is maximized at a given confidence level. To see why this is true, consider plotting $p \times (1 - p)$ versus p . This can be done by typing

```
> p <- seq(0, 1, 0.001)
> plot(p, p*(1-p), type="l")
```

Consequently, solving (8.46) for n yields 4268.4. To guarantee the maximum margin of error is within 0.015 at a 95% confidence level, always take the ceiling of n (use the next largest integer). In this case, a sample of size 4269 will guarantee the maximum margin of error will be less than 0.015 at a 95% confidence level. That is,

$$(1.96)\sqrt{\frac{(\frac{1}{2})(\frac{1}{2})}{4269}} = 0.01499902 < 0.015. \quad \blacksquare$$

8.4.2 Confidence Interval for a Difference in Population Proportions

In this section, the focus is on two populations, X and Y , from which random samples of sizes n_X and n_Y , respectively, are taken. If π_X and π_Y are the population proportions of successes and P_X and P_Y are the respective sample proportions of successes, then the resulting sampling distributions of P_X and P_Y , provided n_X and n_Y are sufficiently large, are approximately normal. That is,

$$P_X \dot{\sim} N\left(\pi_X, \sqrt{\frac{\pi_X(1-\pi_X)}{n_X}}\right) \quad \text{and} \quad P_Y \dot{\sim} N\left(\pi_Y, \sqrt{\frac{\pi_Y(1-\pi_Y)}{n_Y}}\right).$$

Since the sampling distributions of both P_X and P_Y are approximately normal, the sampling distribution for the difference between P_X and P_Y will also be approximately normal. Specifically,

$$P_X - P_Y \dot{\sim} N\left(\pi_X - \pi_Y, \sqrt{\frac{\pi_X(1-\pi_X)}{n_X} + \frac{\pi_Y(1-\pi_Y)}{n_Y}}\right) \quad (8.47)$$

according to Theorem 5.1 on page 176.

Using a similar approach to the one presented for the construction of a confidence interval for the difference between two means, construct a $(1 - \alpha) \cdot 100\%$ asymptotic confidence interval for $\pi_X - \pi_Y$ as shown in (8.48). The rationale for replacing π_X and π_Y with p_X and p_Y in $\sqrt{\frac{\pi_X(1-\pi_X)}{n_X} + \frac{\pi_Y(1-\pi_Y)}{n_Y}}$ is the invariance property of maximum likelihood estimators (property 2 on page 273), where $\hat{\pi}_X = P_X$ and $\hat{\pi}_Y = P_Y$.

$$CI_{1-\alpha}(\pi_X - \pi_Y) = \left[\begin{array}{l} (p_X - p_Y) - z_{1-\alpha/2} \sqrt{\frac{p_X(1-p_X)}{n_X} + \frac{p_Y(1-p_Y)}{n_Y}}, \\ (p_X - p_Y) + z_{1-\alpha/2} \sqrt{\frac{p_X(1-p_X)}{n_X} + \frac{p_Y(1-p_Y)}{n_Y}} \end{array} \right], \quad (8.48)$$

It is generally advisable to use the continuity correction $\frac{1}{2} \left(\frac{1}{n_X} + \frac{1}{n_Y} \right)$ with (8.48) anytime

$$|p_X - p_Y| > \frac{1}{2} \left(\frac{1}{n_X} + \frac{1}{n_Y} \right). \quad (8.49)$$

The continuity correction is subtracted and added to the lower and upper confidence limits of (8.48), respectively. The S function `prop.test()` automatically applies the continuity correction when (8.49) is satisfied provided the user does not issue the argument `correct=FALSE`.

Example 8.26 A company wants to see if a certain change in the process for manufacturing component parts is beneficial. Samples are taken using both the existing and the new procedure to determine if the new process results in an improvement. The first sample is taken before the change has been implemented, and the second sample is taken once the change has been implemented. If 70 of 1400 elements are found to be defective in the first sample and 90 of 2000 elements are found to be defective from the second sample, find a 95% confidence interval for the true difference in the proportion of defective components between the existing and the new processes.

Solution: The sample proportions of successes are $p_X = \frac{70}{1400} = 0.05$ and $p_Y = \frac{90}{2000} = 0.045$. Using (8.48), the 95% confidence interval for the true difference in the proportion of defective components between the existing and the new processes is given in (8.50). Since the confidence interval contains 0, there is no reason to suspect the new procedure significantly reduces the proportion of defective items.

$$\begin{aligned} CI_{0.95}(\pi_X - \pi_Y) &= \\ & \left[\begin{array}{l} (0.05 - 0.045) - 1.96 \sqrt{\frac{(0.05)(1-0.05)}{1400} + \frac{(0.045)(1-0.045)}{2000}}, \\ (0.05 - 0.045) + 1.96 \sqrt{\frac{(0.05)(1-0.05)}{1400} + \frac{(0.045)(1-0.045)}{2000}} \end{array} \right] \\ &= [-0.0096, 0.0196] \end{aligned} \quad (8.50)$$

To construct a 95% confidence interval for $\pi_X - \pi_Y$ using (8.48), key in

```
> round(prop.test(c(70,90), c(1400,2000), correct=FALSE)$conf,4)
[1] -0.0096 0.0196
attr(,"conf.level")
[1] 0.95
```

Since $|p_X - p_Y| = |0.05 - 0.045| = 0.005 > \frac{1}{2} \left(\frac{1}{1400} + \frac{1}{2000} \right) = 0.0006$, 0.0006 should be subtracted from and added to the smaller and larger values reported in (8.50), respectively. Consequently, a continuity corrected 95% confidence interval for the true difference in the proportion of defective components between the existing and the new process is

$$CI_{0.95}(\pi_X - \pi_Y) = [-0.0096 - 0.0006, 0.0196 + 0.0006] = [-0.0102, 0.0202].$$

To produce the continuity corrected interval with S, enter

```
> round(prop.test(c(70,90), c(1400,2000), correct=TRUE)$conf,4)
[1] -0.0102 0.0202
attr(,"conf.level")
[1] 0.95
```

■

8.4.3 Confidence Interval for the Mean of a Poisson Random Variable

Recall that a Poisson random variable counts the number of occurrences over some period of time or region of space where the occurrences are relatively rare. When collecting occurrences from a Poisson distribution, it follows that the sample values will have a positive skew, since the Poisson distribution itself is skewed to the right. This will often rule out confidence interval formulas that require normality assumptions. However, for sufficiently large samples, one can use (8.36) on page 322 to construct confidence limits for the mean of a Poisson distribution. When using (8.36) for confidence interval construction for the mean of a Poisson random variable, first find the maximum likelihood estimator of λ . In Example 7.18 on page 262, the maximum likelihood estimator of λ for a Poisson distribution was found to be \bar{X} . That is, $\hat{\lambda}(\mathbf{X}) = \bar{X}$. To calculate the Fisher information $I_n(\lambda)$ using (7.49) on page 270 requires knowledge of the second-order partial derivative of the log-likelihood function with respect to λ . This second-order partial derivative was computed in Example 7.18 and is reproduced here for the reader's benefit:

$$\frac{\partial^2 \ln L(\lambda|\mathbf{X})}{\partial \lambda^2} = \frac{-\sum_{i=1}^n x_i}{\lambda^2} \quad (8.51)$$

Taking the expected value of (8.51) yields the following, from which the Fisher information, $I_n(\lambda)^{-1} = \frac{\lambda}{n}$, is obtained:

$$-E \left[\frac{\partial^2 \ln L(\lambda|\mathbf{X})}{\partial \lambda^2} \right] = \frac{n\lambda}{\lambda^2} = \frac{n}{\lambda}. \quad (8.52)$$

Taking advantage of the asymptotic properties of MLE estimators allows one to write

$$\hat{\lambda}(\mathbf{X}) = \bar{X} \sim N \left(\lambda, \sqrt{\frac{\lambda}{n}} \right) \text{ as } n \rightarrow \infty.$$

One may then use (8.36), the confidence interval formula for MLEs, to construct a $(1 - \alpha) \cdot 100\%$ asymptotic confidence interval for λ as shown here where $\hat{\sigma}_{\hat{\lambda}(\mathbf{x})} = \sqrt{\frac{\bar{x}}{n}}$:

$$CI_{1-\alpha}(\lambda) = \left[\bar{x} - z_{1-\alpha/2} \sqrt{\frac{\bar{x}}{n}}, \bar{x} + z_{1-\alpha/2} \sqrt{\frac{\bar{x}}{n}} \right] \quad (8.53)$$

One could obtain a similar confidence interval by recognizing that \bar{X} has a normal distribution with parameters μ and $\frac{\sigma}{\sqrt{n}}$ for large sample sizes according to the Central Limit Theorem. Since the mean for a Poisson is λ and the standard deviation of a Poisson random variable is $\sqrt{\lambda}$, it follows that

$$\bar{X}_{Pois} \sim N\left(\lambda, \frac{\sqrt{\lambda}}{\sqrt{n}}\right).$$

Example 8.27 Example 4.4 on page 122 provided evidence to suggest the number of goals scored in the regulation 90 minute periods of World Cup soccer matches from 1990 to 2002 have a Poisson distribution. Use the information in column `Goals` of the data set `Soccer` to construct a 90% confidence interval for the mean number of goals scored during a 90 minute regulation period.

Solution: The 90% confidence interval for λ is constructed using (8.53):

$$\begin{aligned} CI_{0.90}(\lambda) &= \left[\bar{x} - z_{1-0.10/2} \sqrt{\frac{\bar{x}}{n}}, \bar{x} + z_{1-0.10/2} \sqrt{\frac{\bar{x}}{n}} \right] \\ &= \left[2.48 - 1.645 \sqrt{\frac{2.48}{232}}, 2.48 + 1.645 \sqrt{\frac{2.48}{232}} \right] \\ &= [2.31, 2.65] \end{aligned} \tag{8.54}$$

To compute the values in (8.54) with `S`, attach `Soccer` and key in

```
> M<-mean(Goals, na.rm=TRUE)
> M
[1] 2.478448
> z<-qnorm(.95)
> z
[1] 1.644854
> round(c(M-z*sqrt(M/232), M+z*sqrt(M/232)), 2)
[1] 2.31 2.65
```

So, one is 90% confident the mean number of goals scored in a World Cup soccer match lies in [2.31, 2.65]. ■

8.5 Problems

1. Is $[\bar{x} - 3, \bar{x} + 3]$ a confidence interval for the population mean of a normal distribution? Why or why not?
2. Explain how to construct a confidence interval for the population mean of a normal distribution with a 95% confidence level.
3. Given a random sample $\{X_1, X_2, \dots, X_n\}$ from a normal population $N(\mu, \sigma)$, where σ is known:
 - (a) What is the confidence level for the interval $\bar{x} \pm 1.881 \frac{\sigma}{\sqrt{n}}$?
 - (b) What is the confidence level for the interval $\bar{x} \pm 1.175 \frac{\sigma}{\sqrt{n}}$?
 - (c) What is the value of the percentile $z_{\alpha/2}$ for a 92% confidence interval?
4. Given a random sample $\{X_1, X_2, \dots, X_n\}$ from a normal population $N(\mu, \sigma)$, where σ is known, consider the confidence interval $\bar{x} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$ for μ .
 - (a) Given a fixed sample size n , explain the relationship between the confidence level and the precision of the confidence interval.
 - (b) Given a confidence level $(1-\alpha)\%$, explain how the precision of the confidence interval changes with the sample size.
5. Given a normal population with known variance σ^2 , by what factor must the sample size be increased to reduce the length of a confidence interval for the mean by a factor of k ?
6. A historic data set studied by R.A. Fisher is the measurements in centimeters of four flower parts (sepal length, sepal width, petal length, and petal width) on 50 specimens for each of three species of irises (*Setosa*, *Versicolor*, and *Virginica*). The data are named **iris** in S-PLUS, and the same data can be found in R under the name **iris3** (Fisher, 1936).
 - (a) Analyze the sepal lengths for *Setosa*, *Versicolor*, and *Virginica* irises, and comment on the characteristics of their distributions. (Hint: Since the data in **iris** and **iris3** are stored as arrays, type `iris3[,1,1]` if using R or `iris[,1,1]` if using S-PLUS to isolate the sepal lengths for the *Setosa* irises.)
 - (b) Based on the analysis from part (a), construct an appropriate 95% confidence interval for the mean sepal length of *Setosa* irises.
7. Surface-water salinity measurements were taken in a bottom-sampling project in White-water Bay, Florida. These data are stored in the data frame **Salinity** in the PASWR package. Geographic considerations lead geologists to believe that the salinity variation should be normally distributed. If this is true, it means there is free mixing and interchange between open marine water and fresh water entering the bay (Davis, 1986).
 - (a) Construct a quantile-quantile plot of the data. Does this plot rule out normality?
 - (b) Construct a 90% confidence interval for the mean salinity variation.

8. The survival times in weeks for 20 male rats that were exposed to a high level of radiation are

152 152 115 109 137 88 94 77 160 165
125 40 128 123 136 101 62 153 83 69

Data are from Lawless (1982) and are stored in the data frame **Rat**.

- (a) Construct a quantile-quantile plot of the survival times. Based on the quantile-quantile plot, can normality be ruled out?
- (b) Construct a 97% confidence interval for the average survival time for male rats exposed to high levels of radiation.
9. A school psychologist administered the Stanford-Binet intelligence quotient (IQ) test in two counties. Forty randomly selected gifted and talented students were selected from each county. The Stanford-Binet IQ test is said to follow a normal distribution with a mean of 100 and standard deviation of 16. The data collected are stored in the data frame **SBIQ**.

County1							County2						
130	126	139	126	124	149	124	127	125	127	132	139	132	125
138	138	140	127	140	124	124	130	131	140	130	132	134	128
121	125	134	121	125	126	122	137	121	121	141	141	137	126
137	146	127	124	142	122	126	124	124	128	145	123	126	132
124	126	121	138	124	126	137	135	126	128	144	121	135	125
122	131	128	122	144			125	136	122	130	130		

- (a) Although the standard deviation for the Stanford-Binet IQ test is known, should it be used? Justify.
- (b) Be careful, the confidence interval formula that should be used in this situation has not been explicitly covered yet. Construct a 90% confidence interval for the true average IQ difference for gifted and talented students between the two counties.
10. A large company wants to estimate the proportion of its accounts that are paid on time.
- (a) How large a sample is needed to estimate the true proportion within 2% with a 95% confidence level?
- (b) Suppose 650 out of 800 accounts are paid on time. Construct a 99% confidence interval for the true proportion of accounts that paid on time.
11. In a study conducted at Appalachian State University, students used digital oral thermometers to record their temperatures each day they came to class. A randomly selected day of student temperatures is provided in the following table and in the data frame **StatTemps**. Information is also provided with regard to subject gender and the hour of the day when the students' temperatures were measured.

	8 a.m. Class			9 a.m. Class		
Males	92.7	94.1	96.5	94.1	96.0	98.2
	93.2	97.1	93.7	96.5	94.4	
Females	96.9	94.0	93.7	96.5	94.3	93.9
	93.9	93.5	97.0	96.5	95.6	98.2
	97.2	92.0	96.6	96.4	96.3	95.1
	94.9	92.1		97.1	96.6	96.8

- (a) Construct a 95% confidence interval for the true average temperature difference between males and females. Does the interval contain the value zero? What does this suggest about gender temperature differences?
- (b) Construct a 95% confidence interval for the true average temperature difference between students taking their temperatures at 8 a.m. and students taking their temperatures at 9 a.m. Give a reason why one group appears to have a higher temperature reading.
12. The Cosmed K4b² is a portable metabolic system. A study at Appalachian State University compared the metabolic values obtained from the Cosmed K4b² to those of a reference unit (Amatek) over a range of workloads from easy to maximal to test the validity and reliability of the Cosmed K4b². A small portion of the results for VO₂ (ml/kg/min) measurements taken at a 150 watt workload are stored in data frame **CosAma** and in the following table:

Subject	Cosmed	Amatek	Subject	Cosmed	Amatek
1	31.71	31.20	8	30.33	27.95
2	33.96	29.15	9	30.78	29.08
3	30.03	27.88	10	30.78	28.74
4	24.42	22.79	11	31.84	28.75
5	29.07	27.00	12	22.80	20.20
6	28.42	28.09	13	28.99	29.25
7	31.90	32.66	14	30.80	29.13

- (a) Construct a quantile-quantile plot for the between system differences.
- (b) Are the VO₂ values reported for Cosmed and Amatek independent?
- (c) Construct a 95% confidence interval for the average VO₂ system difference.
13. Let $\{X_1, \dots, X_9\}$ and $\{Y_1, \dots, Y_{15}\}$ be two random samples from a $N(\mu_X, \sigma)$ and a $N(\mu_Y, \sigma)$, respectively. Suppose that $\bar{x} = 57.3$, $s_X^2 = 8.3$, $\bar{y} = 65.6$, and $s_Y^2 = 9.7$. Find a 96% confidence interval for μ_X , μ_Y , and $\mu_X - \mu_Y$.
14. The water consumption in liters per family per day in a given city is a normally distributed random variable with unknown variance. Consider the following confidence intervals for the population mean obtained from a random sample of size n :

$$[374.209, 545.791], \quad [340.926, 579.074], \quad [389.548, 530.452].$$

- (a) Find the value of the sample mean.
- (b) If the intervals are obtained from the same random sample, match the confidence levels 90%, 95% and 99% with the corresponding confidence intervals.
15. The best-paid 20 tennis players in the world have earned millions of dollars during their careers and are famous for having won some of the four “Grand Slam” tournaments. Somewhat less famous players who are in positions 20 through 100 in the earnings’ rankings have also garnered large sums. The following data (in millions of dollars) correspond to the earnings of 15 randomly selected players classified somewhere in positions 20 through 100. They are also stored in the data frame **Top20**.

10.10 8.80 8.64 7.67 6.34 6.03 5.90 5.68
5.51 5.38 5.31 4.92 4.54 4.02 3.86

Compute a 94% confidence interval for the average earnings of players classified between positions 20 and 100 of the ranking. (Source: <http://www.atptennis.com/en/>)

16. The following data is the amount of nuclear energy (in TOE, tons of oil equivalent) produced in 12 randomly selected European countries during 2003. The values are also stored in the data frame **TOE**.

12222 6674 15961 3994 2841 1036
1343 4608 5864 17390 22877 4457

Compute a 95% confidence interval for the 2003 mean European TOE assuming the amount of nuclear energy is normally distributed.

17. A group of engineers working with physicians in a research hospital is developing a new device to measure blood glucose levels. Based on measurements taken from patients in a previous study, the physicians assert that the new device provides blood glucose levels slightly higher than those provided by the old device. To corroborate their suspicion, 15 diabetic patients were randomly selected, and their blood glucose levels were measured with both the new and the old devices. The measurements, in mg/100 ml, appear in the following table and are stored in the data frame **glucose**:

Blood glucose levels

Patient	Old	New	Patient	Old	New
Patient 1	182.47	195.64	Patient 9	179.04	195.25
Patient 2	175.53	196.31	Patient 10	180.50	194.48
Patient 3	181.71	190.33	Patient 11	182.15	197.33
Patient 4	179.03	192.90	Patient 12	183.55	193.81
Patient 5	177.28	193.24	Patient 13	180.86	198.03
Patient 6	177.49	193.05	Patient 14	180.82	193.31
Patient 7	179.54	193.87	Patient 15	178.88	198.43
Patient 8	185.12	196.39			

- (a) Are the samples independent? Why or why not?

- (b) If the blood glucose level is a normally distributed random variable, compute a 95% confidence interval for the difference of the population means.
- (c) Use the results in (b) to decide whether or not the two devices give the same results.
18. The European Union is developing new policies to promote research and development investment. A random sample of 15 countries' investments for the years 2002 and 2003 is taken and the results (in millions of euros) are stored in the data frame **EURD** and shown in the following table:

Country	2002	2003
Belgium	5200.737	5177.444
Czech Republic	959.362	1012.579
Estonia	55.699	66.864
France	34527.000	34569.095
Cyprus	33.791	40.969
Latvia	41.532	37.724
Lithuania	99.642	110.580
Hungary	705.754	693.057
Malta	11.861	11.453
Portugal	1029.010	1019.580
Slovenia	360.419	377.435
Slovakia	148.335	169.105
Bulgaria	81.228	88.769
Croatia	270.606	291.856
Romania	183.686	202.941

- (a) Compute a 95% confidence interval for the difference between 2002 and 2003 investment means.
- (b) Use (a) to decide if the new policies are increasing investments.
19. The “Wisconsin Card Sorting Test” is widely used by psychiatrists, neurologists, and neuropsychologists with patients who have a brain injury, neurodegenerative disease, or a mental illness such as schizophrenia. Patients with any sort of frontal lobe lesion generally do poorly on the test. The data frame **WCST** and the following table contain the test scores from a group of 50 patients from the *Virgen del Camino* Hospital (Pamplona, Spain).

23	12	31	8	19	11	36	94	6	10	22	7	18	26	35	78	11
7	28	25	17	8	20	47	5	13	28	19	7	19	38	8	15	40
19	42	17	6	8	6	11	10	19	65	13	17	5	26	15	4	

- (a) Use the function `EDA()` from the **PASWR** package to explore the data and decide if normality can be assumed.

- (b) What assumption(s) must be made to compute a 95% confidence interval for the population mean?
- (c) Compute the confidence interval from (b).
20. The following data were taken to measure the unknown pH values μ of a solution in a chemical experiment:

8.01, 8.05, 7.96, 8.04, 8.03, 8.03, 8.02, 7.98, 8.05, 8.03.

If the pH meter has a systematic error, Δ , and a normally distributed random error, $\varepsilon \sim N(0, \sigma^2)$, then it can be assumed that the observations come from a normal random variable, $X \sim N(\mu + \Delta, \sigma^2)$.

- (a) Compute a 95% confidence interval for μ when $\Delta = 0$ and $\sigma = 0.05$. Compute the interval assuming that the variance is unknown.
- (b) Repeat part (a) with $\Delta = 0.2$.
21. When sampling from a normal distribution, what sample size will ensure that the interval $\bar{x} \pm s$ attains at least a 95% confidence level?
22. Let $\{X_1, \dots, X_n\}$ be a simple random sample from a normal distribution $N(\mu, \sigma)$, and consider the following random variables:

$$X = \min_{1 \leq i \leq n} \{x_i\}, \quad Y = \max_{1 \leq i \leq n} \{x_i\}.$$

- (a) Set the seed value at 69, and generate $m = 100$ samples of size $n = 5$ from a normal population $N(\mu = 5, \sigma = 2)$. Compute the number of intervals of the types $[X, Y]$ containing the real value $\mu = 5$. If the theoretical coverage of these intervals is 94% for a sample of size $n = 5$, do the empirical results agree with the theoretical coverage?
- (b) Set the seed value at 18, and generate $m = 100$ samples of size $n = 5$ from a normal population $N(\mu = 5, \sigma = 2)$. Compute the confidence intervals of the type $[X, Y]$ and $[\bar{X} + z_{0.03} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{0.97} \frac{\sigma}{\sqrt{n}}]$. Construct a plot with the length of both types of intervals. Repeat the exercise with samples of size $n = 50$. Which type of confidence interval is preferred? Why?
23. Given the following data

25.3 23.8 27.5 23.2 24.5 25.3 24.6 26.8 25.9 29.2,

- (a) State the assumption(s) needed to construct a confidence interval for the population variance.
- (b) Assuming your assumption(s) in (a) are satisfied, construct a 95% confidence interval for σ .
- (c) Assuming that $\mu = 25$, construct a 95% confidence interval for σ .
24. Schizophrenia is believed to cause changes in dopamine levels. Twenty-five patients with schizophrenia were classified as psychotic or non-psychotic after being treated with an antipsychotic drug. Samples of cerebral fluid were taken from each patient and assayed for dopamine *b*-hydroxylase (DBH) activity. The dopamine measurements for the two

groups are in nmol/(ml)(h)/(mg) of protein and are stored in the data frame **Schizo** as well as in the following table (Sternberg et al., 1982).

Judged Non-Psychotic			Judged Psychotic	
0.0104	0.0105	0.0112	0.0150	0.0204
0.0116	0.0130	0.0145	0.0208	0.0222
0.0154	0.0156	0.0170	0.0226	0.0245
0.0180	0.0200	0.0200	0.0270	0.0275
0.0210	0.0230	0.0252	0.0306	0.0320

- Construct side-by-side boxplots of the two groups. Based on the boxplots, comment on the relative shapes of the two distributions.
 - Construct quantile-quantile plots for the two groups, and comment on whether or not the plots support the analysis in part (a).
 - Construct a 95% confidence interval for the true ratio of psychotic to non-psychotic variances.
 - Based on the confidence interval for the ratio of variances, should the variances be pooled to construct a 95% confidence interval for the true dopamine level difference between psychotic and non-psychotic patients?
 - Construct a 95% confidence interval for the true dopamine level difference between psychotic and non-psychotic patients.
 - Does the confidence interval contain zero? What does this say about the effectiveness of the antipsychotic drug?
25. Assuming two independent random samples of sizes 22 and 45 with variance estimates of $s_1^2 = 38.7$ and $s_2^2 = 45.6$, respectively, have been taken, construct a 95% confidence interval for σ .
26. Those teams who win Formula 1 championships have pit crews who change tires as fast as possible. The data frame **Formula1** and the following table contain the times (in seconds) that the pit crews of two different teams spent changing tires in 10 randomly selected races.

Race	1	2	3	4	5	6	7	8	9	10
Team 1	5.613	6.130	5.422	5.947	5.514	5.322	5.690	5.243	5.920	5.859
Team 2	5.934	5.335	5.826	4.821	5.664	5.292	5.257	6.245	5.981	5.197

- Assuming that the times are normally distributed, compute a 95% confidence interval for the variance ratio σ_1^2/σ_2^2 . Are the population variances equal?
 - Use the results in part (a) to compute a 95% confidence interval for the difference of the population means $\mu_2 - \mu_1$. What does the result mean?
27. Let $\{X_1, X_2, \dots, X_n\}$ be a random sample from a normal population $N(\mu, \sigma)$, where μ and σ are unknown. Find the value of the sample size n if $(0.59s^2, 2s^2)$ is to be at least a 94% confidence interval for σ^2 .

28. Use a seed equal to 55, and simulate $m = 100$ samples of size $n = 800$ from a $N(15, \sigma = \sqrt{6})$. Calculate the confidence intervals for σ^2 at the $1 - \alpha = 0.96$ confidence level. Plot the confidence intervals, and calculate the number of times the parameter is not contained in the simulated confidence intervals.
29. Use a seed equal to 224, and simulate $m_x = 100$ samples of size $n_x = 1500$ from a $N(3, \sigma = \sqrt{5})$ and $m_y = 100$ samples of size $n_y = 1500$ from a $N(6, \sigma = \sqrt{7})$. Calculate the confidence intervals for σ_x^2/σ_y^2 with a $1 - \alpha = 0.94$ confidence level. Plot the intervals and calculate the number of times the parameter ratio is not in the simulated confidence interval.
30. The drug Sulphinpyrazole was studied for its efficacy in preventing death after myocardial infarction. Construct a 90% confidence interval for the true proportion of deaths between patients who have suffered a myocardial infarction who were administered Sulphinpyrazole and patients who were administered a placebo after myocardial infarctions. Based on the confidence interval, does Sulphinpyrazole appear to reduce the proportion of deaths among patients who have suffered a myocardial infarction?

	Death (all causes)	Survivors
Sulphinpyrazole	41	692
Placebo	60	682

31. From a random sample of 2000 Internet domains registered in a country during the last few years, 300 were “.org” domains. Compute a 98% confidence interval for the proportion of “.org” domains registered in that country during the last few years.
32. Use a seed equal to 10, and simulate 300 samples of size $n_x = 65$ from a $N(4, \sigma_x = \sqrt{2})$ distribution and 300 samples of size $n_y = 90$ from a $N(5, \sigma_y = \sqrt{3})$. Check that $\frac{s_x^2/\sigma_x^2}{s_y^2/\sigma_y^2}$ follows an $F_{64,89}$ distribution.
33. Use a seed equal to 95, and simulate $m = 500$ samples of size $n = 1000$ from a $B(1, \pi = 0.4)$ distribution. Show that the sampling proportion is normally distributed.
34. How large a sample is needed to ensure the bound on the error of estimation for the population proportion is no more than 2 percentage points for a 95% confidence interval?
35. A large company wants to estimate the proportion of its accounts that are paid on time.
- How large a sample is needed to estimate the true proportion within 5% with a 90% confidence interval?
 - Suppose 650 out of 800 accounts are paid on time. Construct a 99% confidence interval for the true proportion of accounts that are paid on time.
36. A sociology research center conducts a survey to discern whether the proportion of vegetarians is larger in urban or rural areas. Of the 180 people from urban areas, 32 were vegetarians. Of the 75 from rural areas, 17 were vegetarians. Construct a 98% confidence interval for the difference between urban and rural vegetarian proportions.
37. Schizophrenia and other psychoses are complex and debilitating diseases, which affect about 2% of the population. Two of the approaches used, as well as in other medical diseases, to reduce clinical heterogeneity among psychoses are categorical and dimensional. The first one assumes that there exist different subgroups within psychosis and

the second one assumes that schizophrenia dimensions fall on a dimensional continuum within psychosis. A sample of 660 consecutively admitted patients in Hospital Virgen del Camino (Pamplona, Spain) is available with the following diagnoses: 358 schizophrenic patients, 61 with schizophreniform disorder, 37 with schizoaffective disorder, 64 with bipolar disorder, 24 with delusional disorder, 54 with brief psychotic disorder, and 32 with atypical psychosis. Compute a 95% confidence interval for the proportion of the different types of patients (Cuesta et al., 2007).

Chapter 9

Hypothesis Testing

9.1 Introduction

A **hypothesis test** in the Neyman-Pearson paradigm is a decision criterion that allows practitioners of statistics to select between two complementary hypotheses. Before conducting the hypothesis test, define the **null hypothesis**, H_0 , which is assumed to be true prior to conducting the hypothesis test. The null hypothesis is compared to another hypothesis, called the **alternative hypothesis**, and denoted H_1 . The alternative hypothesis is often called the research hypothesis since the theory or what is believed to be true about the parameter is specified in the alternative hypothesis. Both hypotheses define complementary subsets of the parameter space Θ where the parameter θ is defined. The null hypothesis defines the region $[\theta \in \Theta_0]$ and the alternative hypothesis defines the region $[\theta \in \Theta_1]$. The subsets Θ_0 and Θ_1 are mutually exclusive by definition, and they are complementary since $\Theta_0 \cup \Theta_1 = \Theta$. When a hypothesis uniquely specifies the distribution of the population from which the sample is taken, the hypothesis is said to be **simple**. For a simple hypothesis, Θ_0 is composed of a single element. Any hypothesis that is not a simple hypothesis is called a **composite hypothesis**. A composite hypothesis does not completely specify the population distribution. Of the various combinations of hypotheses that could be examined, the case where the null hypothesis is simple and the alternative hypothesis is composite will be the focus of this text. Hypothesis tests will generally take a form similar to those in Table 9.1, where θ_0 is a single numerical value. For alternative hypotheses (A) and (B), which are lower one-sided and upper one-sided, respectively, the hypothesis test is called a **one-tailed test**. For the alternative hypothesis in (C), a two-sided alternative, the hypothesis test is called a **two-tailed test**.

Table 9.1: Form of hypothesis tests

Null Hypothesis	Alternative Hypothesis	Type of Alternative
$H_0 : \theta = \theta_0$	(A) $H_1 : \theta < \theta_0$	lower one-sided
	(B) $H_1 : \theta > \theta_0$	upper one-sided
	(C) $H_1 : \theta \neq \theta_0$	two-sided

Example 9.1 If $H_0 : \pi = 0.4$ in a *Bernoulli*(π) distribution, the null hypothesis is simple since the hypothesis $H_0 : \pi = 0.4$ uniquely specifies the distribution as *Bernoulli*(0.4). If $H_1 : \pi < 0.4$, the hypothesis is composite since π can take any value in the interval $[0, 0.4)$.

The goal in hypothesis testing is to decide which one of the two hypotheses (null and alternative) is true. To this end, split the sample space into two mutually exclusive subsets R and R^c . R is the **rejection region** and R^c is referred to as the **acceptance region**. The

critical value is the number that splits Θ into R and R^c . To help decide between the two hypotheses, calculate a test statistic based on the sample information from the experiment. If the test statistic falls in the acceptance region, accept the null hypothesis. If the value of the test statistic falls in the rejection region, reject the null hypothesis and accept the alternative hypothesis.

There are two basic ways to think of a hypothesis test. First, one can think of it as a two-decision problem where the researcher will choose one of two hypotheses to be true. This is the historical approach due to Jerzy Neyman and Egon Pearson. The second method, due to Ronald Fisher, determines how much evidence exists in the data against the null hypothesis. The null hypothesis is never accepted but is merely a hypothesis of “no difference.” The test will determine if the data that have been collected could be due to chance alone if the null hypothesis were true; and if this is not likely, the researcher has statistically significant evidence that the alternative hypothesis is true. A hypothesis test where the null hypothesis is never accepted but merely “not rejected” is called a significance test.

Example 9.2 The weight of a ball-bearing fluctuates between 1.5 g and 4.5 g. One wants to test whether the distribution of the weight for the ball-bearing has a mean of either 2 g ($H_0 : \mu = 2$) or 2.5 g ($H_1 : \mu = 2.5$). A random sample of size one is taken. If the weight of the ball-bearing is greater than 2.3 g, the null hypothesis that the mean weight of the ball-bearings is 2 g is rejected, and the alternative hypothesis that the mean weight of the ball-bearings is 2.5 g is accepted. Specify the sample space, the rejection region, and the acceptance region for this experiment.

Solution: The sample space is given by the interval $[1.5, 4.5]$. The rejection region is the subinterval $R = (2.3, 4.5]$, and the acceptance region is the subinterval $R^c = [1.5, 2.3]$. Note that $R^c \cup R = [1.5, 2.3] \cup (2.3, 4.5] = [1.5, 4.5]$. ■

9.2 Type I and Type II Errors

The decision one reaches using a hypothesis test is always subject to error. That is, when a decision is reached to reject the null hypothesis and accept the alternative hypothesis, this may be the correct decision or a mistake (error). Likewise, if the null hypothesis is not rejected but rather accepted, an error could also be made. Simply put, one can never be sure of the truth since the decision in a hypothesis test to reject or not to reject a hypothesis is based on sample information. To get a better grasp on the errors one might make with a hypothesis test, consider the following hypothetical legal situation.

An individual is on trial for a capital offense. In the United States’ judicial system, an individual is considered innocent until proven guilty of an offense. Consequently, the null hypothesis in this case is that the individual is innocent and the alternative hypothesis is that the person is guilty. After the prosecuting and defending attorneys present their evidence, the jury makes a decision either to convict or not to convict the individual of the capital offense. If the prosecuting attorney presents a strong case, the jury is likely to convict the defendant. However, just because the jury convicts the defendant, it does not mean that the defendant is actually guilty of the capital offense. Likewise, if the jury does not convict the defendant of the capital offense, this does not imply the individual is innocent. To better see the possible consequences of the decisions the jury may reach, consider Table 9.2 on the facing page.

Table 9.2: Possible outcomes and their consequences for a trial by jury

	True State of the Defendant (Reality)	
	H_0 True (innocent)	H_0 False (guilty)
Jury's Decision		
Accept H_0 (not guilty)	A. correct	B. error
Reject H_0 (guilty)	C. error	D. correct

- A. If the null hypothesis is true and the null hypothesis is accepted, the decision is correct. In the legal example, if the defendant is innocent and the jury decides the defendant is not guilty of the charge, the jury's decision is correct.
- B. If the null hypothesis is false and it is not rejected, the decision is incorrect. By failing to reject a false null hypothesis, an error has been made. In statistics, this error is called a **type II error**. The probability of committing a type II error is β . In the legal scenario, a type II error is made when a guilty person is not convicted.
- C. If the null hypothesis is true and it is rejected, the decision is incorrect. In other words, by rejecting a true null hypothesis, an error has been made. In statistics, this type of error is called a **type I error**. The probability of committing a type I error is α . In the legal example, a type I error would be to convict an innocent defendant.
- D. If the null hypothesis is false and it is rejected, the decision is correct. In the legal arena, this translates into a jury convicting a guilty defendant.

The probability of committing a type I error (rejecting H_0 when it is true) is called the **level of significance** for a hypothesis test. The level of significance is also known as the size of the test and is denoted by α , where

$$\alpha = \mathbb{P}(\text{type I error}) = \mathbb{P}(\text{reject } H_0 | H_0 \text{ is true}) = \mathbb{P}(\text{accept } H_1 | H_0 \text{ is true}).$$

The probability of committing a type II error is β , where

$$\begin{aligned} \beta &= \mathbb{P}(\text{type II error}) = \mathbb{P}(\text{fail to reject } H_0 | H_0 \text{ is false}) \\ &= \mathbb{P}(\text{accept } H_0 | H_1 \text{ is true}). \end{aligned}$$

The relationship between type I and type II errors is shown in Table 9.3 on the next page.

If the researcher fails to reject the null hypothesis when the null hypothesis is true, note that no error is committed. Specifically, the correct decision should be reached in roughly $(1 - \alpha) \times 100\%$ of all trials. Using the same logic, approximately $(1 - \beta) \times 100\%$ of the times sample data are evaluated in a test of hypothesis, a false null hypothesis will be rejected.

Since a type I error is frequently considered to be more serious than a type II error and the probability of a type I error is easier to control than the probability of a type II error, it is common practice for researchers to specify *a priori* the largest probability of a type I error they are willing to accept and subsequently to use this value as their level of significance to make a decision when they conduct their hypothesis testing. The North American judicial system certainly considers convicting an innocent person to be a worse error than allowing a guilty person to walk free. However, a type I error is not always more critical than a type II error. Suppose one is going to go sky diving. In this scenario, the null hypothesis

Table 9.3: Relationship between type I and type II errors

		Decision	
		Reject H_0	Fail To Reject H_0
Null Hypothesis	True	Type I Error $\mathbb{P}(\text{Type I Error}) = \alpha$ (Level of Significance)	Correct Decision $\mathbb{P}(\text{Accept } H_0 H_0) = 1 - \alpha$
	False	Correct Decision $\mathbb{P}(\text{Accept } H_1 H_1) = 1 - \beta$ (Power of the Test)	Type II Error $\mathbb{P}(\text{Type II Error}) = \beta$

is that the parachute will open and the alternative hypothesis is that the parachute will not open. Certainly a type II error (concluding the parachute will open when it will not) is more critical than a type I error (concluding the parachute will not open when it will).

Example 9.3 Given a normal distribution with unknown mean μ and known standard deviation $\sigma = 2$, one wishes to test the null hypothesis $H_0 : \mu = 1$ versus the alternative hypothesis $H_1 : \mu = 4$. A sample of size one is taken where the rejection region is considered to be the interval $(2, \infty)$. In other words, if the sample value is greater than 2, the null hypothesis is rejected. On the other hand, if the sample value is less than or equal to two, one fails to reject the null hypothesis. Determine α and β for this experiment.

Solution: Although there is no way to know if the decision made with regard to the null hypothesis is correct, there is a reasonable criterion that allows the determination of the probability of making type I and type II errors.

Determine α — The probability of committing a type I error, the level of significance, is the probability that the sample value falls in the rejection region $(2, \infty)$ when $H_0 : \mu = 1$ is true. To find α , it is necessary to find $\mathbb{P}(X_1 > 2|N(1, 2))$. See Figure 9.1 for a graphical representation of the type I error. Note that

$$\alpha = \mathbb{P}(X_1 > 2|N(1, 2)) = \mathbb{P}\left(\frac{X_1 - 1}{2} > \frac{2 - 1}{2}\right) = \mathbb{P}(Z > 0.5) = 0.31.$$

To find α with S, key in

```
> ALPHA <- round(1 - pnorm(2, 1, 2), 2)
> ALPHA
[1] 0.31
```

Note: S-PLUS returns the area to the left of a given value when using the function `pnorm`. By default, R also returns the area to the left of a given value when using the function `pnorm`. However, R also allows the user to find the area to the right of a given value by using the argument `lower.tail=FALSE`. Consequently, one might have used the `lower.tail=FALSE` argument with R's `pnorm` function to find the answer.

```
> ALPHA <- round(pnorm(2, 1, 2, lower.tail=FALSE), 2) # Only with R
> ALPHA
[1] 0.31
```

Determine β — The probability of making a type II error is the probability of failing to reject $H_0 : \mu = 1$ when in actuality $H_1 : \mu = 4$. In other words, although $\mu = 4$, the null hypothesis is not rejected because the test statistic does not fall in the rejection region but does lie in the region $(-\infty, 2]$. For a graphical representation of the type II error, see Figure 9.1. Mathematically this is written

$$\beta = \mathbb{P}(X_1 \leq 2 | N(4, 2)) = \mathbb{P}(Z \leq -1) = 0.16.$$

To find β with S, enter

```
> BETA <- round(pnorm(2,4,2), 2)
> BETA
[1] 0.16
```

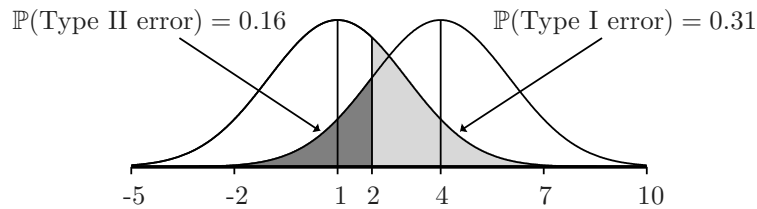


FIGURE 9.1: Graphical representation of type I and type II errors when $H_0 : \mu = 1$ versus $H_1 : \mu = 4$. ■

Since the probabilities of committing type I and type II errors for a fixed sample size are dependent, it is usually impossible to make both type I and type II errors arbitrarily small. However, out of convenience, the tests considered are restricted to only those tests that control the type I error at a given significance level and subsequently select from these tests the test with the most power. Researchers typically fix the probability of committing a type I error at the 0.01, 0.05, or 0.1 significance level; however, these are merely values that were tabled early in the history of statistics and have been used mainly for convenience rather than through any actual merit. Since there are as many tests as there are partitions of the sample space, the number of tests one may have to evaluate to decide between two competing hypotheses might be huge. For this very reason, certain partitions will produce results that are more appealing in the sense of supporting a specific hypothesis.

9.3 Power Function

Given a composite alternative hypothesis $H_1 : \theta \in \Theta_1$, the power of the test, $Power(\theta)$, is

$$\begin{aligned} Power(\theta) &= \mathbb{P}(\text{reject } H_0 | H_0 \text{ is false}) = \mathbb{P}(\text{accept } H_1 | H_1) \\ &= 1 - \beta(\theta), \end{aligned} \tag{9.1}$$

where $\beta(\theta)$ is the probability of a type II error at a given θ . Loosely speaking, the power of a test is the probability the test detects differences when differences exist. Note that

$Power(\theta)$ is a function of the parameter θ , which has for each value of θ in the alternative hypothesis, $\theta \in \Theta_1$, the power that a simple alternative hypothesis would have for that value of θ . When the null hypothesis is simple, $\theta = \theta_0$, the power of the test at θ_0 is the same as the significance level, that is, $Power(\theta_0) = \alpha$.

Example 9.4 Given the density function

$$f(x; \theta) = \theta e^{-\theta x}, \quad x \geq 0, \quad \theta > 0,$$

- (a) Consider a test of hypothesis where $H_0 : \theta = 2$ versus $H_1 : \theta > 2$. Using a random sample of size one, find k such that the test is conducted at the $\alpha = 0.05$ level.
- (b) Further, determine the power function of this test.

Solution: The solutions are as follows:

- (a) First, set up the integral to find the value of k that yields a significance level of 0.05:

$$\alpha = \mathbb{P}(X_1 < k | H_0) = \int_0^k 2e^{-2x_1} dx_1 = 1 - e^{-2k} = 0.05$$

The solution for k is $k = 0.02564665$.

```
> qexp(0.05, 2)
[1] 0.02564665
```

- (b) The power of the test is

$$\begin{aligned} Power(\theta) &= 1 - \beta(\theta) = \mathbb{P}(\text{accept } H_1 \text{ when it is true}) = \mathbb{P}(X_1 < 0.0256 | H_1) \\ &= \int_0^{0.0256} \theta e^{-\theta x_1} dx_1 = 1 - e^{-0.0256\theta}. \end{aligned}$$

Note that the answer clearly illustrates that it is not possible to obtain a single value for the power of a composite alternative hypothesis since the answer itself is a function of θ . In other words, for each value of the parameter θ compatible with the alternative hypothesis (in this case $\theta > 2$), a value for the power function is obtained that corresponds to that simple hypothesis. As the parameter θ takes on values greater than two, the power function approaches one. ■

Example 9.5 \triangleright **Achievement Test** \triangleleft Test the null hypothesis that for a certain age group the mean score on an achievement test (scores follow a normal distribution with $\sigma = 6$) is equal to 40 against the alternative that it is not equal to 40.

- (a) Find the probability of type I error for $n = 9$ if the null hypothesis is rejected when the sample mean is less than 36 or greater than 44.
- (b) Find the probability of type I error for $n = 36$ if the null hypothesis is rejected when the sample mean is less than 38 or greater than 42.
- (c) Plot the power functions for $n = 9$ and $n = 36$ for values of μ between 30 and 50.

Solution: The solutions are as follows:

(a) The probability of a type I error for $n = 9$ if the null hypothesis is rejected when the sample mean is less than 36 or greater than 44 is

$$\begin{aligned}\mathbb{P}(\text{Type I error}) &= \mathbb{P}\left(\bar{X} < 36 \mid N\left(40, \frac{6}{\sqrt{9}}\right)\right) + \mathbb{P}\left(\bar{X} > 44 \mid N\left(40, \frac{6}{\sqrt{9}}\right)\right) \\ &= \mathbb{P}\left(Z < \frac{36 - 40}{2}\right) + \mathbb{P}\left(Z > \frac{44 - 40}{2}\right) \\ &= \mathbb{P}(Z < -2) + \mathbb{P}(Z > 2) = 0.02275 + 0.02275 = 0.04550.\end{aligned}$$

To compute the answer with S, key in

```
> pnorm(36,40,6/sqrt(9)) + 1 - pnorm(44,40,6/sqrt(9))
[1] 0.04550026
```

(b) The probability of type I error for $n = 36$ if the null hypothesis is rejected when the sample mean is less than 38 or greater than 42 is

$$\begin{aligned}\mathbb{P}(\text{Type I error}) &= \mathbb{P}\left(\bar{X} < 38 \mid N\left(40, \frac{6}{\sqrt{36}}\right)\right) + \mathbb{P}\left(\bar{X} > 42 \mid N\left(40, \frac{6}{\sqrt{36}}\right)\right) \\ &= \mathbb{P}\left(Z < \frac{38 - 40}{1}\right) + \mathbb{P}\left(Z > \frac{42 - 40}{1}\right) \\ &= \mathbb{P}(Z < -2) + \mathbb{P}(Z > 2) = 0.02275 + 0.02275 = 0.04550.\end{aligned}$$

To compute the answer with S, enter

```
> pnorm(38,40,6/sqrt(36)) + 1 - pnorm(42,40,6/sqrt(36))
[1] 0.04550026
```

(c) The power function for $n = 9$ is

$$\text{Power}(\mu) = \mathbb{P}\left(\bar{X} < 36 \mid N\left(\mu, \frac{6}{\sqrt{9}}\right)\right) + \mathbb{P}\left(\bar{X} > 44 \mid N\left(\mu, \frac{6}{\sqrt{9}}\right)\right)$$

The power function for $n = 36$ is

$$\text{Power}(\mu) = \mathbb{P}\left(\bar{X} < 38 \mid N\left(\mu, \frac{6}{\sqrt{36}}\right)\right) + \mathbb{P}\left(\bar{X} > 42 \mid N\left(\mu, \frac{6}{\sqrt{36}}\right)\right)$$

To produce a plot similar to the one in Figure 9.2 on the next page with R, use the following code:

```
> mu <- seq(30,50,.01)
> power9 <- 1-pnorm(44, mu,6/sqrt(9)) + pnorm(36, mu,6/sqrt(9))
> power36 <- 1-pnorm(42, mu,6/sqrt(36)) + pnorm(38, mu,6/sqrt(36))
> plot(mu,power9,type="l", ylab=expression(Power(mu)), xlab=expression(mu),
+ ylim=c(0,1))
> lines(mu, power36, type="l")
> arrows(32, 0.6 , 34.2, .78, lwd=2, length=0.05)
> arrows(32, 0.35 , 37, .78, lwd=2, length=0.05)
> arrows(40, 0.4 , 40, 0.06, lwd=2, length=0.05)
> text(32,0.58, expression(n==9))
> text(32.3,0.33, expression(n==36))
> text(40,0.45, expression(alpha==0.045))
```

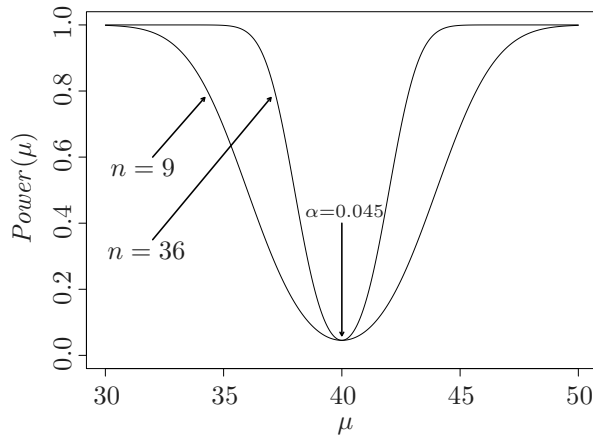


FIGURE 9.2: Graphical representation of the power function, $Power(\mu)$, for both scenarios in Example 9.5 on page 346.

Note that $Power(\mu_0) = \alpha$ for both power functions depicted in Figure 9.2. In general, as the true μ is farther from the hypothesized μ in H_0 , the power of a test will increase. Additionally, the power function approaches 1 faster for larger n as the true μ moves farther from the hypothesized μ in H_0 . ■

9.4 Uniformly Most Powerful Test

First, note that tests with identical α values do not necessarily have identical power for a fixed sample size as in Example 9.6.

Example 9.6 Given a $N(\mu, 1)$ population from which one takes a simple random sample of size 1, test the null hypothesis $H_0 : \mu = 1$ versus the alternative hypothesis $H_1 : \mu = 2$. Determine the significance level and the power of the test for the following rejection regions:

- (a) $(2.036, \infty)$
- (b) $(1.100, 1.300) \cup (2.461, \infty)$.

Solution: The answers are as follows:

- (a) Since $R = (2.036, \infty)$,

$$\alpha = \mathbb{P}(X > 2.036 | N(1, 1)) = \mathbb{P}\left(\frac{X - 1}{1} > \frac{2.036 - 1}{1}\right) = \mathbb{P}(Z > 1.036) = 0.150,$$

$$\beta = \mathbb{P}(X \leq 2.036 | N(2, 1)) = \mathbb{P}\left(\frac{X - 2}{1} \leq \frac{2.036 - 2}{1}\right) = \mathbb{P}(Z \leq 0.036) = 0.514,$$

and the power of the test is $1 - \beta = 1 - 0.514 = 0.486$. See Figure 9.3 on the facing page for a graphical representation of the type I and type II errors.

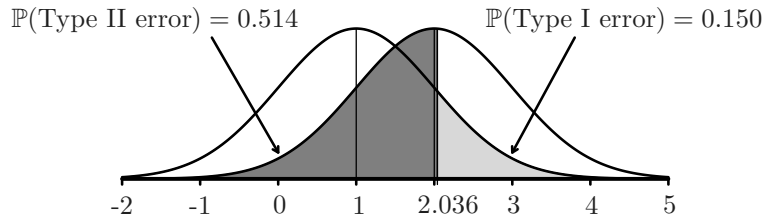


FIGURE 9.3: Graphical representation of type I and type II errors when $H_0 : \mu = 1$ versus $H_1 : \mu = 2$ with rejection region $(2.036, \infty)$.

(b) Since the rejection region is $(1.100, 1.300) \cup (2.461, \infty)$, the probability of committing a type I error is

$$\begin{aligned}
 \alpha &= \mathbb{P}(1.100 < X < 1.300 | N(1, 1)) + \mathbb{P}(X > 2.461 | N(1, 1)) \\
 &= \mathbb{P}\left(\frac{1.100 - 1}{1} < Z < \frac{1.300 - 1}{1}\right) + \mathbb{P}\left(\frac{X - 1}{1} > \frac{2.461 - 1}{1}\right) \\
 &= \mathbb{P}(0.100 < Z < 0.300) + \mathbb{P}(Z > 1.461) \\
 &= \mathbb{P}(Z < 0.300) - \mathbb{P}(Z < 0.100) + \mathbb{P}(Z > 1.461) = 0.618 - 0.540 + 0.072 = 0.150,
 \end{aligned}$$

and the probability of committing a type II error is

$$\begin{aligned}
 \beta &= \mathbb{P}(X \leq 1.100 | N(2, 1)) + \mathbb{P}(1.300 \leq X \leq 2.461 | N(2, 1)) \\
 &= \mathbb{P}\left(\frac{X - 2}{1} \leq \frac{1.100 - 2}{1}\right) + \mathbb{P}\left(\frac{1.300 - 2}{1} \leq \frac{X - 2}{1} \leq \frac{2.461 - 2}{1}\right) \\
 &= \mathbb{P}(Z \leq -0.900) + \mathbb{P}(-0.700 \leq Z \leq 0.461) \\
 &= \mathbb{P}(Z \leq -0.900) + \mathbb{P}(Z \leq 0.461) - \mathbb{P}(Z \leq -0.700) \\
 &= 0.184 + 0.678 - 0.242 = 0.620.
 \end{aligned}$$

It follows that the power of the test is $1 - \beta = 1 - 0.620 = 0.380$. A graphical representation of the type I and type II errors is provided in Figure 9.4 on the next page. To find α and β with **S**, type

```

> ALPHA <- pnorm(1.300, 1, 1) - pnorm(1.100, 1, 1) + (1 - pnorm(2.461, 1, 1))
> round(ALPHA, 3)
[1] 0.15
> BETA <- pnorm(1.100, 2, 1) + pnorm(2.461, 2, 1) - pnorm(1.300, 2, 1)
> round(BETA, 3)
[1] 0.62

```

■

It is clear to see from the previous example that, with the same level of significance (0.150), the power obtained for the test with a rejection region of $(1.100, 1.300) \cup (2.461, \infty)$ has less power than the test that uses a rejection region of $(2.036, \infty)$. The probabilities of committing type I and type II errors for the rejection regions $(2.036, \infty)$ and $(1.100, 1.300) \cup (2.461, \infty)$ are shown in Figures 9.3 and 9.4, respectively. In general, it is possible to have a

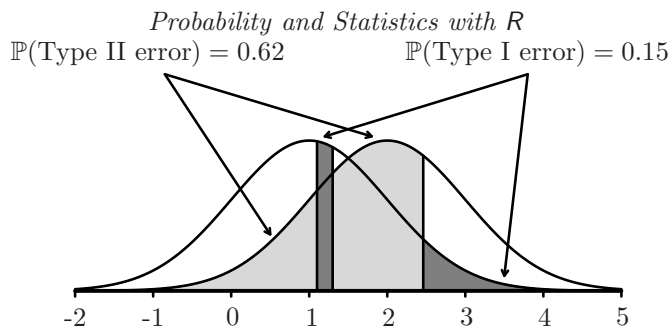


FIGURE 9.4: Graphical representation of type I and type II errors when $H_0 : \mu = 1$ versus $H_1 : \mu = 2$ with rejection region $(1.100, 1.300) \cup (2.461, \infty)$

test that is “better” in the sense of having more power than another test even though both tests have the same significance level. So, the researcher wants to find a **uniformly most powerful** test that has more power than all other tests that have the correct significance level, α , if such a test exists. To be complete, it is important to note that uniformly most powerful tests do not always exist. A generalization that can be made from Example 9.6 is that one-sided tests with the same sample size as two-sided tests will always have more power for the same α level.

9.5 \wp -Value or Critical Level

Fisher’s advocates object to establishing *a priori* the level of significance when testing a hypothesis. Instead, they prefer to make their decisions to reject or fail to reject the null hypothesis based on \wp -values. The critical level or **\wp -value** is defined as the probability of observing a difference as extreme or more extreme than the difference observed under the assumption that the null hypothesis is true. Virtually all statistical software packages will return a \wp -value when testing a hypothesis. The values of the statistic $t(\mathbf{x})$ observed from the sample and \wp -value calculations are summarized in Table 9.4.

Table 9.4: Calculation of \wp -values for continuous distributions

\wp -Value
$H_1 : \theta < \theta_0$ $\mathbb{P}(T \leq t_{\text{obs}} H_0)$
$H_1 : \theta > \theta_0$ $\mathbb{P}(T \geq t_{\text{obs}} H_0)$
$H_1 : \theta \neq \theta_0$ $2 \min\left\{\mathbb{P}(T \leq t_{\text{obs}} H_0), \mathbb{P}(T \geq t_{\text{obs}} H_0)\right\}$

It is important to note that the \wp -value is not fixed *a priori*, but rather is determined after the sample is taken. A small \wp -value indicates that observing differences as large or larger than the one found in the sample is rare, and thus do not occur by chance alone. A small \wp -value lends support to H_1 ; so, given a fixed significance level α , reject H_0 whenever the \wp -value $< \alpha$. In Fisher’s paradigm, hypothesis tests are tests of significance, where

a φ -value is calculated without regard to a fixed rejection region. The Neyman-Pearson paradigm uses a specified α level to calculate a rejection region that is used in conjunction with a standardized test statistic to reach a statistical conclusion.

9.6 Tests of Significance

Using the following steps incorporates ideas from both Fisher and Neyman and Pearson for solving **test of hypothesis**-type problems. The steps allow others to follow the reasoning one uses to reach a statistical decision.

Step 1: Hypotheses — State the null and alternative hypotheses.

First, establish the null hypothesis, $H_0 : \theta = \theta_0$. Next, determine the form of the alternative hypothesis, H_1 . The forms H_1 can take are found in Table 9.1 on page 341, where evidence is to be found that θ is less than, greater than, or not equal to the θ_0 specified in H_0 . If one wishes to specify a value for which H_1 is true, that value is denoted with either θ_1 or $\theta_1(X, Y, \dots)$.

Step 2: Test Statistic — Select an appropriate test statistic and determine the sampling distribution of the test statistic or the standardized test statistic under the assumption that the null hypothesis is true.

Choose a test statistic, $\hat{\theta}$, generally one such that the expected value of the test statistic is equal to the parameter in H_0 . For example, if testing μ , $\hat{\theta} = \bar{X}$; or, if testing π , $\hat{\theta} = P$.

A common standardized test statistic will take the form

$$T = t(\mathbf{X}) = \frac{\hat{\theta}(\mathbf{X}) - \theta_0}{\sqrt{\text{Var}[\hat{\theta}(\mathbf{X})]}}$$

Other test statistics will present themselves when testing hypotheses regarding variances.

Step 3: Rejection Region Calculations — If the computations are to be done by hand, use the specified α level to compute the critical value and to determine the rejection region for the standardized test statistic. If the computations are to be done by a computer, do not do this.

Then, calculate the value of $t(\mathbf{X})$, assuming H_0 is true. The value of the statistic $t(\mathbf{X})$ observed from the sample is denoted $t(\mathbf{x}) = t_{\text{obs}}$.

Step 4: Statistical Conclusion — If a rejection region was not computed in step 3, calculate the φ -value. The procedure for calculating the φ -value is found in Section 9.5 on the facing page.

Use the rejection region or the φ -value to determine if the evidence warrants rejecting the null hypothesis. If t_{obs} falls into the rejection region, reject H_0 ; if not, fail to reject H_0 . If the φ -value is less than α , reject H_0 ; if not, fail to reject H_0 .

Step 5: **English Conclusion** — State in plain English what the conclusion reached in step 4 means. This statement will always be about the alternative hypothesis. That is, the evidence will either warrant concluding the alternative hypothesis or the evidence will not be sufficient to conclude the alternative hypothesis is true.

There are two distributions that occur frequently in hypothesis testing involving means: a standard normal distribution and a t -distribution. When the standardized test statistic follows a standard normal distribution, the hypothesis test will typically be called a **one-sample z -test** or a **two-sample z -test**, depending on whether there are one or two samples. Likewise, if the standardized test statistic follows a t -distribution, the test will be a **one-sample t -test**, a **two-sample t -test**, or a **paired t -test**. The general form for a z -test statistic is

$$\frac{\text{statistic} - \mu_{\text{statistic}}}{\sigma_{\text{statistic}}} \quad (9.2)$$

while the general form of a t -test statistic is

$$\frac{\text{statistic} - \mu_{\text{statistic}}}{\hat{\sigma}_{\text{statistic}}}. \quad (9.3)$$

Duality of Confidence Intervals and Tests of Significance When confidence intervals were constructed in Chapter 8, there was often a statistic $\hat{\theta}(\mathbf{X})$ that had a known distribution, where θ was the mean of $\hat{\theta}(\mathbf{X})$ and $\sigma_{\hat{\theta}(\mathbf{X})}$ was the square root of the variance of $\hat{\theta}(\mathbf{X})$. From this statistic's distribution, a pivot was constructed that took the form $\frac{\hat{\theta}(\mathbf{X}) - \theta}{\sigma_{\hat{\theta}(\mathbf{X})}}$ with a known distribution (denoted, in general, as T). One would use this pivot to construct a $(1 - \alpha) \cdot 100\%$ confidence interval:

$$CI_{1-\alpha}(\theta) = \left[\hat{\theta}(\mathbf{x}) + t_{\alpha/2} \cdot \sigma_{\hat{\theta}(\mathbf{x})}, \hat{\theta}(\mathbf{x}) + t_{1-\alpha/2} \cdot \sigma_{\hat{\theta}(\mathbf{x})} \right].$$

In testing hypotheses, when the standardized test statistic has the same form as the pivot used to construct a confidence interval, namely $t_{\text{obs}} = \frac{\hat{\theta}(\mathbf{x}) - \theta_0}{\sigma_{\hat{\theta}(\mathbf{x})}}$, and the confidence intervals and the acceptance region for the null hypothesis are based on the same distribution, there exists a duality between $(1 - \alpha) \cdot 100\%$ confidence intervals and α -level hypothesis tests. That is, when θ_0 is in the confidence interval, $H_0 : \theta = \theta_0$ is not rejected. This is summarized in general in Table 9.5.

Table 9.5: Duality of $(1 - \alpha) \cdot 100\%$ confidence intervals and α -level tests of significance

Alternative Hypothesis	Fail to Reject H_0 Region	$(1 - \alpha) \cdot 100\%$ Confidence Interval
$H_1 : \theta < \theta_0$	$t_{\text{obs}} \geq t_{\alpha}$	$\left(-\infty, \hat{\theta}(\mathbf{x}) - t_{\alpha} \cdot \sigma_{\hat{\theta}(\mathbf{x})} \right]$
$H_1 : \theta > \theta_0$	$t_{\text{obs}} \leq t_{1-\alpha}$	$\left[\hat{\theta}(\mathbf{x}) - t_{1-\alpha} \cdot \sigma_{\hat{\theta}(\mathbf{x})}, \infty \right)$
$H_1 : \theta \neq \theta_0$	$t_{\alpha/2} \leq t_{\text{obs}} \leq t_{1-\alpha/2}$	$\left[\hat{\theta}(\mathbf{x}) + t_{\alpha/2} \cdot \sigma_{\hat{\theta}(\mathbf{x})}, \hat{\theta}(\mathbf{x}) + t_{1-\alpha/2} \cdot \sigma_{\hat{\theta}(\mathbf{x})} \right]$

9.7 Hypothesis Tests for Population Means

9.7.1 Test for the Population Mean when Sampling from a Normal Distribution with Known Population Variance

The null hypothesis for testing the mean when sampling from a normal distribution with known variance is $H_0 : \mu = \mu_0$, where μ_0 is a particular value. It is important to emphasize that a normal distribution as well as a known variance are being assumed. Seldom, if ever, will the distribution and its variance be known with certainty. However, a firm foundation in how significance tests are conducted with these assumptions will provide a foundation on which more hypothesis testing procedures can be built.

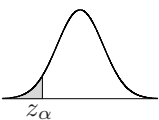
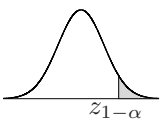
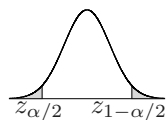
The basic idea behind a test of significance for the mean when working with a random sample of size n is to determine how likely the values observed in the sample are to occur. Typically, the sampling distribution of \bar{X} , which is $N(\mu_0, \sigma/\sqrt{n})$, is used to construct a standardized test statistic since one is sampling from a normal distribution under the assumption that the null hypothesis is true. Further, the Central Limit Theorem states that the sampling distribution of \bar{X} approaches a normal distribution even if the original population is not normal, provided the sample size n is sufficiently large. The standardized test statistic under the assumption that H_0 is true is

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1).$$

The formula to calculate its observed value as well as the three possible alternative hypotheses and their rejection regions are described in Table 9.6.

Table 9.6: Summary for testing the mean when sampling from a normal distribution with known variance (one-sample z -test)

Null Hypothesis — $H_0 : \mu = \mu_0$ Standardized Test Statistic's Value — $z_{\text{obs}} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$

Alternative Hypothesis	$H_1 : \mu < \mu_0$	$H_1 : \mu > \mu_0$	$H_1 : \mu \neq \mu_0$
Rejection Region	$z_{\text{obs}} < z_\alpha$	$z_{\text{obs}} > z_{1-\alpha}$	$ z_{\text{obs}} > z_{1-\alpha/2}$
Graphical Representation of Rejection Region			

Example 9.7 A random sample of size $n = 30$ is taken from a distribution known to be $N(\mu, \sigma = 2)$. If the $\sum_{i=1}^{30} x_i = 56$,

- (a) Test the null hypothesis $H_0 : \mu = 1.8$ versus the alternative hypothesis $H_1 : \mu > 1.8$ at the $\alpha = 0.05$ significance level.

(b) Find $\beta(3)$ and $Power(3)$.

Solution: The answers are as follows:

(a) Use the five-step procedure.

Step 1: **Hypotheses** — $H_0 : \mu = 1.8$ versus $H_1 : \mu > 1.8$.

Step 2: **Test Statistic** — The test statistic chosen is \bar{X} because $E[\bar{X}] = \mu$. The value of this test statistic is $\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{56}{30} = 1.867$. The standardized test statistic and its distribution under the assumption H_0 is true are $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$.

Step 3: **Rejection Region Calculations** — Because the standardized test statistic is distributed $N(0, 1)$, and H_1 is an upper one-sided hypothesis, the rejection region is $z_{\text{obs}} > z_{1-\alpha} = z_{0.95} = 1.64$. The value of the standardized test statistic is $z_{\text{obs}} = \frac{1.867 - 1.8}{2/\sqrt{30}} = 0.183$.

Step 4: **Statistical Conclusion** — The φ -value is $\mathbb{P}(Z \geq 0.183) = 0.427$.

- I. From the rejection region, fail to reject H_0 because 0.183 is not greater than 1.64.
- II. From the φ -value, fail to reject H_0 because the φ -value = 0.427 is greater than 0.05.

Fail to reject H_0 .

Step 5: **English Conclusion** — There is not evidence to suggest that the mean is greater than 1.8.

To use S to find $z_{0.95}$, the φ -value for a z_{obs} value of 0.183 for a right tail alternative hypothesis, key in

```
> qnorm(0.95)                # Critical Value
[1] 1.644854
> 1 - pnorm(.183)           # P-value
[1] 0.427399
```

(b) $\beta(3)$ and $Power(3)$ are

$$\begin{aligned}
 \beta(3) &= \mathbb{P}(\text{Type II error}) = \mathbb{P}\left(\text{Fail to reject } H_0 \mid N\left(3, \frac{2}{\sqrt{30}}\right)\right) \\
 &= \mathbb{P}\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha} \mid N\left(3, \frac{2}{\sqrt{30}}\right)\right) \\
 &= \mathbb{P}\left(\bar{X} \leq z_{0.95} \frac{\sigma}{\sqrt{n}} + \mu \mid N\left(3, \frac{2}{\sqrt{30}}\right)\right) \\
 &= \mathbb{P}\left(\frac{\bar{X} - 3}{\frac{2}{\sqrt{30}}} \leq \frac{\frac{(1.645)(2)}{\sqrt{30}} + 1.8 - 3}{\frac{2}{\sqrt{30}}}\right) \\
 &= \mathbb{P}(Z \leq -1.645) = 0.05
 \end{aligned}$$

$$Power(3) = 1 - \beta(3) = 1 - 0.05 = 0.95$$

To use S to find $\beta(3)$ and $Power(3)$, enter

```

> beta3 <- round(pnorm(qnorm(.95,1.8,2/sqrt(30)),3,2/sqrt(30)), 2)
> power3 <- 1 - beta3
> beta3
[1] 0.05
> power3
[1] 0.95

```



9.7.2 Test for the Population Mean when Sampling from a Normal Distribution with Unknown Population Variance

The null hypothesis is still $H_0 : \mu = \mu_0$ when working with data from a normal distribution with unknown variance. However, the standardized test statistic under the assumption that H_0 is true is now

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}.$$

The formula to calculate its observed value as well as the three possible alternative hypotheses and their rejection regions are described in Table 9.7 on the next page. The computation of β and $Power$ with the t -test is not nearly as easy as with the standard normal distribution. This is due to the fact that when the null hypothesis is false, the random variable $\frac{\bar{X} - \mu_0}{S/\sqrt{n}}$ has what is known as a non-central t -distribution with non-centrality parameter


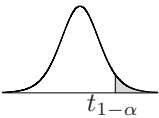
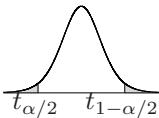
$$\gamma = \frac{\mu_1 - \mu_0}{\frac{\sigma}{\sqrt{n}}},$$

denoted $t_{n-1;\gamma}^*$, where μ_1 is the true value of μ . Currently, S-PLUS does not have a non-central t -distribution. However, S-PLUS does have a non-central F distribution that can be used to compute the power for a two-tailed alternative hypothesis involving the t -test using the relationship in (9.4). R has both a non-central t -distribution and a non-central F

distribution. To compute the power for a t -test, one must provide some estimate of σ for the non-centrality parameter:

$$\mathbb{P}((t_{n-1;\gamma}^* < t_{\alpha/2;n-1}) \cup (t_{n-1;\gamma}^* > t_{1-\alpha/2;n-1})) = \mathbb{P}(F_{1,n-1;\gamma^2} > (t_{1-\alpha/2;n-1})^2) \quad (9.4)$$

Table 9.7: Summary for testing the mean when sampling from a normal distribution with unknown variance (one-sample t -test)

	Null Hypothesis — $H_0 : \mu = \mu_0$	Standardized Test Statistic's Value	— $t_{\text{obs}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
Alternative Hypothesis	$H_1 : \mu < \mu_0$	$H_1 : \mu > \mu_0$	$H_1 : \mu \neq \mu_0$
Rejection Region	$t_{\text{obs}} < t_{\alpha;n-1}$	$t_{\text{obs}} > t_{1-\alpha;n-1}$	$ t_{\text{obs}} > t_{1-\alpha/2;n-1}$
Graphical Representation of Rejection Region			
Note that the degrees of freedom for the t values in all the graphical representations are $n - 1$.			

Example 9.8 A random sample of size $n = 25$ is taken from a distribution known to be $N(\mu, \sigma)$. If the $\sum_{i=1}^n x_i = 100$ and the $\sum_{i=1}^n x_i^2 = 600$,

- Test the null hypothesis $H_0 : \mu = 2.5$ versus the alternative hypothesis $H_1 : \mu \neq 2.5$ at the $\alpha = 0.05$ significance level.
- Find $\text{Power}(\mu_1 = 4)$ if it is assumed $\sigma = 2.5$.
- Use **S** to simulate a $t_{24;\gamma=3}^*$ distribution, and use it to compute the simulated power of the test in (b).

Solution: The answers are as follows:

- To solve this part, use the five-step procedure.

Step 1: **Hypotheses** — These are given in the problem as

$$H_0 : \mu = 2.5 \text{ versus } H_1 : \mu \neq 2.5$$

Step 2: **Test Statistic** — The test statistic chosen is \bar{X} because $E[\bar{X}] = \mu$. The value of this test statistic is $\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{100}{25} = 4$. The standardized test statistic and its distribution under the assumption H_0 is true are $\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{25-1}$.

Step 3: **Rejection Region Calculations** — Because the standardized test statistic is distributed t_{24} , and H_1 is a two-tailed hypothesis, the rejection region is $|t_{\text{obs}}| > t_{1-0.05/2;24} = t_{0.975;24} = 2.06$. The value of the standardized test statistic is $t_{\text{obs}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{4 - 2.5}{2.89/\sqrt{25}} = 2.595156$. (The value for s is calculated $\sqrt{\frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{n-1}} = \sqrt{\frac{600 - (25)(4^2)}{25-1}} = 2.89$.)

Step 4: **Statistical Conclusion** — The φ -value is $2 \cdot \mathbb{P}(t_{24} \geq 2.595) = 0.016$.

I. From the rejection region, reject H_0 because $t_{\text{obs}} = 2.595$ is greater than 2.06.

II. From the φ -value, reject H_0 because the φ -value = 0.016 is less than 0.05.

Reject H_0 .

Step 5: **English Conclusion** — There is evidence to suggest that the mean is not equal to 2.5.

To use S to find $t_{0.975,24}$ and the φ -value for a t_{obs} value of 2.595 for a two-tailed alternative hypothesis, type

```
> qt(0.975,24)                # Critical Value
[1] 2.063899
> round(2*(1 - pt(2.595156,24)), 3) # P-value
[1] 0.016
```

(b) Before computing $Power(\mu_1 = 4)$, first determine the non-centrality parameter:

$$\gamma = \frac{\mu_1 - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{4.0 - 2.5}{\frac{2.5}{\sqrt{25}}} = 3.0.$$

Let $T = t(\mathbf{X}) = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$. Then

$$\begin{aligned} Power(\mu_1 = 4) &= \mathbb{P}(\text{Reject } H_0 | H_1) \\ &= \mathbb{P}\left((T < t_{\alpha/2; n-1}) \cup (T > t_{1-\alpha/2; n-1}) \mid T \sim t_{n-1}^*; \gamma\right) \\ &= \mathbb{P}\left((t_{24;3}^* < t_{0.025;24}) \cup (t_{24;3}^* > t_{0.975;24})\right) \\ &= \mathbb{P}\left((t_{24;3}^* < -2.06) \cup (t_{24;3}^* > 2.06)\right) \\ &= \mathbb{P}\left((t_{24;3}^* < -2.06) + (t_{24;3}^* > 2.06)\right) = 0.82 \end{aligned}$$

A graphical representation of the $Power(\mu_1 = 4)$ is depicted in Figure 9.5 on the following page. Find the $Power(\mu_1 = 4)$ using the non-central t -distribution and the non-central F distribution in R. Note that if one is using S-PLUS, one can only solve the problem using the non-central F distribution. Further, the non-central F distribution cannot be used to find power for directional hypotheses. To find $\mathbb{P}\left((t_{24;3}^* < t_{0.025;24}) \cup (t_{24;3}^* > t_{0.975;24})\right)$ with R, enter

```
> pt(qt(0.025,24),24,3)+(1-pt(qt(0.975,24),24,3))
[1] 0.8207219
```


Using the relationship between t -distributions and F distributions given in (9.4), write

$$\begin{aligned} \mathbb{P}((t_{24;3}^* < t_{0.025;24}) \cup (t_{24;3}^* > t_{0.975;24})) &= \mathbb{P}(F_{1,24;\gamma=3^2} > (t_{1-\alpha/2;n-1})^2) \\ &= \mathbb{P}(F_{1,24;\gamma=3^2} > (t_{0.975;24})^2) \\ &= \mathbb{P}(F_{1,24;\gamma=3^2} > (2.06)^2 = 4.26) \\ &= 0.82. \end{aligned}$$

To find $\mathbb{P}(F_{1,24;9} > 4.26) = 1 - \mathbb{P}(F_{1,24;9} < 4.26)$ with S, key in

```
> 1-pf(qt(.975,24)^2,1,24,9)
[1] 0.8207219
```

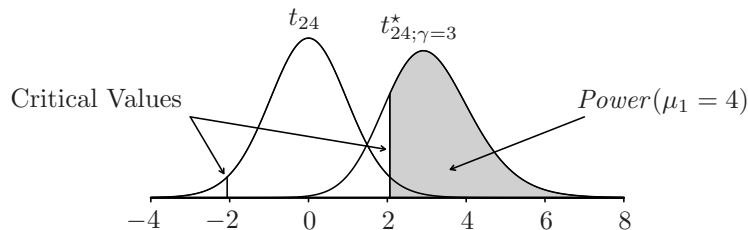


FIGURE 9.5: Central t -distribution and non-central t -distribution with $\gamma = 3$

One can also use the R function `power.t.test()` to compute the answer as follows:

```
> power.t.test(n=25, delta=1.5, sd=2.5, type="one.sample")
```

One-sample t test power calculation

```
      n = 25
  delta = 1.5
      sd = 2.5
sig.level = 0.05
  power = 0.8207213
alternative = two.sided
```

(c) The following S code computes the simulated power and produces the graph in Figure 9.6 on the next page:

```
> set.seed(13)
> nvar <- rnorm(25 * 20000, 4, 2.5)
> nvarmat <- matrix(nvar, 20000, 25) # 20000 by 25 Matrix
> xbar <- apply(nvarmat, 1, mean)
> S <- apply(nvarmat, 1, sd) # Change sd to stdev for S-PLUS
> tstar <- (xbar - 2.5)/(S/5)
> hist(tstar, xlim = c(-4, 8), nclass = "Scott", col = 13,
+ xlab = "", probability = T, ylim = c(0, 0.4),
+ main = "Central and Simulated Non-Central t-Distributions")
> crit.tu <- qt(0.975, 24)
```

```

> crit.tl <- qt(0.025, 24)
> x <- seq(-4, 8, 0.05)
> y <- dt(x, 24)
> lines(x, y, lwd = 2)
> lines(c(-4, 8), c(0, 0), lwd = 3)
> lines(c(crit.tl, crit.tl), c(0, dt(crit.tl, 24)), lwd = 2)
> lines(c(crit.tu, crit.tu), c(0, dt(crit.tu, 24)), lwd = 2)
> SPU <- length(tstar[tstar > crit.tu])/length(tstar)
> SPL <- length(tstar[tstar < crit.tl])/length(tstar)
> Power <- SPU + SPL
> Power
[1] 0.82035

```

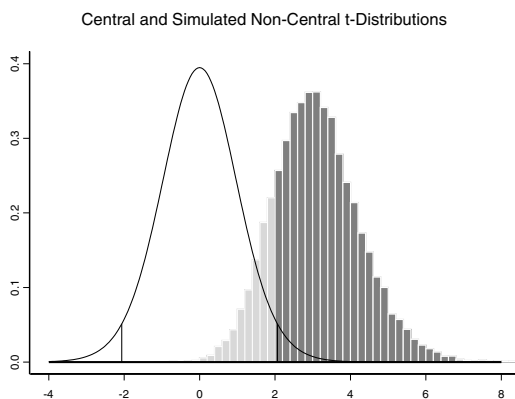


FIGURE 9.6: Central t -distribution and simulated non-central t -distribution with $\gamma = 3$ ■

Example 9.9 ▷ *One-Sample t -Test: Fertilizers* ◁ A farmer wants to test if a new brand of fertilizer increases his wheat yields per plot. He puts the new fertilizer on 15 equal plots and records the subsequent yields for the 15 plots. If his traditional yield is two bushels per plot, conduct a test of significance for μ at the $\alpha = 0.05$ significance level after verifying the data follow a normal distribution. The yields for the 15 fields are

2.5 3.0 3.1 4.0 1.2 5.0 4.1 3.9 3.2 3.3 2.8 4.1 2.7 2.9 3.7

Solution: To solve this problem, start by verifying the normality assumption of the data using exploratory data analysis (`EDA()`). The results from applying the function `EDA()` to the wheat yields per plot are provided in Figure 9.7 on the following page. Based on the graphical output from the function `EDA()`, it is not unreasonable to assume that wheat yield follows a normal distribution. Now, proceed with the five-step procedure.

Step 1: **Hypotheses** — To test if wheat yield is increased, the hypotheses are

$$H_0 : \mu = 2 \text{ versus } H_1 : \mu > 2$$

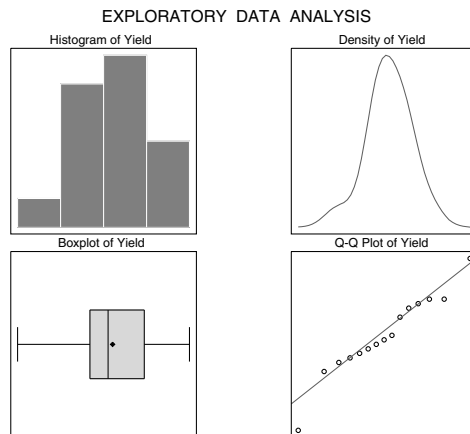


FIGURE 9.7: Exploratory data analysis of the wheat yield per plot values

Step 2: **Test Statistic** — The test statistic chosen is \bar{X} because $E[\bar{X}] = \mu$. The value of this test statistic is $\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{49.5}{15} = 3.3$. The standardized test statistic under the assumption that H_0 is true and its distribution are $\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{15-1}$.

Step 3: **Rejection Region Calculations** — Because the standardized test statistic is distributed t_{14} , and H_1 is an upper one-sided hypothesis, the rejection region is $t_{\text{obs}} > t_{1-0.05;14} = t_{0.95;14} = 1.76$. The value of the standardized test statistic is $t_{\text{obs}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{3.3 - 2}{0.892/\sqrt{15}} = 5.64$.

Step 4: **Statistical Conclusion** — The φ -value is $\mathbb{P}(t_{14} \geq 5.64) \approx 0$.

I. From the rejection region, reject H_0 because $t_{\text{obs}} = 5.64$ is greater than 1.76.

II. From the φ -value, reject H_0 because the φ -value ≈ 0 is less than 0.05.

Reject H_0 .

Step 5: **English Conclusion** — There is evidence to suggest that the mean yield with the new fertilizer is greater than two bushels per plot.

To perform calculations with S, enter

```
> qt(.95,14)                # Critical Value
[1] 1.76131
> x<-c(2.5,3,3.1,4,1.2,5,4.1,3.9,3.2,3.3,2.8,4.1,2.7,2.9,3.7)
> mean(x)
[1] 3.3
> sd(x)                      # for S-PLUS use stdev(x)
[1] 0.8920282
> round(1 - pt(5.64,14),4)   # P-value
[1] 0
```

To compute the value of the standardized test statistic and its corresponding p -value with S, type

```
> t.test(x, alternative="greater", mu=2)
```

```
One Sample t-test
```

```
data: x
t = 5.6443, df = 14, p-value = 3.026e-05
alternative hypothesis: true mean is greater than 2
95 percent confidence interval:
 2.894334      Inf
sample estimates:
mean of x
      3.3
```

Note that the upper limit of the confidence interval in the R output is `Inf` (S-PLUS uses `NA` instead of `Inf`), indicating that the limit on the right side of the confidence interval is ∞ . Also, the calculation of the lower limit uses (8.10) on page 300 modified for a one-sided confidence interval. ■

9.7.3 Test for the Difference in Population Means when Sampling from Independent Normal Distributions with Known Variances

When sampling from two normal distributions with known variances, the null hypothesis for testing the difference between two means is $H_0 : \mu_X - \mu_Y = \delta_0$, and the standardized test statistic under the assumption that H_0 is true is

$$Z = \frac{\bar{X} - \bar{Y} - \delta_0}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \sim N(0, 1).$$

The formulas to calculate its observed value as well as the three possible alternative hypotheses and their rejection regions are described in Table 9.8 on the following page. Note that testing the equality of two means ($H_0 : \mu_X = \mu_Y$) is the same as specifying $\delta_0 = 0$ in the null hypothesis $H_0 : \mu_X - \mu_Y = \delta_0$.

Example 9.10 A researcher wishes to see if it is reasonable to believe that engineering majors have higher math SAT scores than English majors. She takes two random samples. The first sample consists of 64 engineering majors' SAT math scores (X). Typically, these scores follow a normal distribution with a known standard deviation of $\sigma_X = 100$ but with an unknown mean. The second sample consists of 144 observations of English majors' SAT scores (Y). These also follow a normal distribution with a standard deviation of $\sigma_Y = 108$ with an unknown mean as well.

- Test the null hypothesis of equality of means at the 10% significance level ($\alpha = 0.1$) knowing the difference in sample means is 20.
- Find the power of the test in part (a) if $\mu_1(X, Y) = \mu_X - \mu_Y = 40$. (Note that $\mu_0(X, Y) = 0$ from H_0 .)

Table 9.8: Summary for test for differences in means when taking independent samples from normal distributions with known variances (two-sample z -test)Null Hypothesis — $H_0 : \mu_X - \mu_Y = \delta_0$ Standardized
Test Statistic's
Value — $z_{\text{obs}} = \frac{\bar{x} - \bar{y} - \delta_0}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}$

Alternative Hypothesis	Rejection Region
$H_1 : \mu_X - \mu_Y < \delta_0$	$z_{\text{obs}} < z_\alpha$
$H_1 : \mu_X - \mu_Y > \delta_0$	$z_{\text{obs}} > z_{1-\alpha}$
$H_1 : \mu_X - \mu_Y \neq \delta_0$	$ z_{\text{obs}} > z_{1-\alpha/2}$

Solution: The answers are as follows:

(a) Use the five-step procedure.

Step 1: **Hypotheses** — To test if engineering majors have a higher average math SAT score than English majors, the hypotheses are

$$H_0 : \mu_X - \mu_Y = 0 \text{ versus } H_1 : \mu_X - \mu_Y > 0$$

Step 2: **Test Statistic** — The test statistic chosen is $\bar{X} - \bar{Y}$ because $E[\bar{X} - \bar{Y}] = \mu_X - \mu_Y$. The value of this test statistic is 20 according to the problem. The standardized test statistic under the assumption that H_0 is true and its distribution are

$$\frac{\bar{X} - \bar{Y} - \delta_0}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \sim N(0, 1).$$

Step 3: **Rejection Region Calculations** — Because the standardized test statistic is distributed $N(0, 1)$, and H_1 is an upper one-sided hypothesis, the rejection region is $z_{\text{obs}} > z_{1-0.1} = z_{0.9} = 1.28$. The value of the standardized test statistic is

$$z_{\text{obs}} = \frac{\bar{x} - \bar{y} - \delta_0}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} = \frac{20 - 0}{\sqrt{\frac{100^2}{64} + \frac{108^2}{144}}} = \frac{20}{15.403} = 1.2985$$

Step 4: **Statistical Conclusion** — The φ -value is $\mathbb{P}(Z \geq 1.2985) = 0.0971$.I. From the rejection region, reject H_0 because $z_{\text{obs}} = 1.2985$ is greater than 1.28.II. From the φ -value, reject H_0 because the φ -value = 0.097 is less than 0.1.**Reject H_0 .**Step 5: **English Conclusion** — There is evidence to suggest that the difference between the average math SAT score for engineering majors and that of the average math SAT score for English majors is greater than zero; therefore, the evidence suggests engineering majors have a higher average math SAT score.

(b) Find the power of the test in part (a) if

$$H_1 : \bar{X} - \bar{Y} \sim N(\mu_1(X, Y) = 40, \sigma_{\bar{X}-\bar{Y}}).$$

Recall that $\mu_0(X, Y) = 0$ in H_0 and $\sigma_{\bar{X}-\bar{Y}} = \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} = 15.403$.

$$\begin{aligned} \beta(\mu_1(X, Y)) &= \mathbb{P}(\text{Fail to Reject } H_0 | H_1) \\ &= \mathbb{P}\left(\frac{\bar{X} - \bar{Y} - \mu_0(X, Y)}{\sigma_{\bar{X}-\bar{Y}}} \leq z_{0.9} | H_1\right) \\ &= \mathbb{P}\left(\frac{\bar{X} - \bar{Y} - 0}{\sigma_{\bar{X}-\bar{Y}}} \leq z_{0.9} | H_1\right) \\ &= \mathbb{P}(\bar{X} - \bar{Y} \leq z_{0.9} \sigma_{\bar{X}-\bar{Y}} | H_1) \\ &= \mathbb{P}\left(\frac{\bar{X} - \bar{Y} - \mu_1(X, Y)}{\sigma_{\bar{X}-\bar{Y}}} \leq \frac{z_{0.9} \sigma_{\bar{X}-\bar{Y}} - \mu_1(X, Y)}{\sigma_{\bar{X}-\bar{Y}}} | H_1\right) \\ &= \mathbb{P}\left(Z \leq \frac{(1.282)(15.403) - 40}{15.403}\right) \\ &= \mathbb{P}(Z \leq -1.315) = 0.094 \end{aligned}$$

So, the power is

$$\text{Power}(\mu_1(X, Y)) = 1 - \beta(\mu_1(X, Y)) = 1 - 0.094 = 0.906.$$

To find the power with S, enter

```
> sig <- sqrt(100^2/64 + 108^2/144)
> sig
[1] 15.40292
> CV <- qnorm(0.90,0, sig)          # Critical value for xbar - ybar
> CV
[1] 19.73589
> BETA <- pnorm(CV,40, sig)
> BETA
[1] 0.09411202
> POWER <- 1 - BETA
> POWER
[1] 0.905888
```

■

9.7.4 Test for the Difference in Means when Sampling from Independent Normal Distributions with Variances That Are Unknown but Assumed Equal

Recall that when random samples of size n_X and n_Y , respectively, are taken from two normal distributions $N(\mu_X, \sigma)$ and $N(\mu_Y, \sigma)$, where σ is unknown, the random variable

$$T = \frac{[(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)]}{\sqrt{S_p^2 \left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}} \sim t_{n_X+n_Y-2}$$

by Theorem 6.4 on page 237, where $S_p^2 = \frac{(n_X-1)S_X^2 + (n_Y-1)S_Y^2}{n_X+n_Y-2}$. The null hypothesis used to test for a difference of means between two normal distributions where the variances are assumed to be unknown but equal is $H_0 : \mu_X - \mu_Y = \delta_0$. When H_0 is false, the random variable T has a non-central t -distribution with non-centrality parameter

$$\gamma = \frac{\mu_1(X, Y) - \mu_0(X, Y)}{\sigma_{\bar{X}-\bar{Y}}}$$

where $\mu_1(X, Y)$ is the value of $\mu_X - \mu_Y$ under H_1 and $\mu_0(X, Y) = \delta_0$.

This distribution is denoted $t_{n_X+n_Y-2; \gamma}^*$. The value of the standardized test statistic is written

$$t_{\text{obs}} = \frac{\bar{x} - \bar{y} - \delta_0}{\sqrt{\frac{(n_X-1)s_X^2 + (n_Y-1)s_Y^2}{n_X+n_Y-2}} \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} = \frac{\bar{x} - \bar{y} - \delta_0}{s_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}$$

The three possible alternative hypotheses and the corresponding rejection regions are in Table 9.9.

Table 9.9: Summary for test for differences in means when taking independent samples from normal distributions with unknown but assumed equal variances (two-sample pooled t -test)

Null Hypothesis — $H_0 : \mu_X - \mu_Y = \delta_0$

Standardized Test
Statistic's Value — $t_{\text{obs}} = \frac{\bar{x} - \bar{y} - \delta_0}{s_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}$

Alternative Hypothesis	Rejection Region
$H_1 : \mu_X - \mu_Y < \delta_0$	$t_{\text{obs}} < t_{\alpha; n_X+n_Y-2}$
$H_1 : \mu_X - \mu_Y > \delta_0$	$t_{\text{obs}} > t_{1-\alpha; n_X+n_Y-2}$
$H_1 : \mu_X - \mu_Y \neq \delta_0$	$ t_{\text{obs}} > t_{1-\alpha/2; n_X+n_Y-2}$

Use of the pooled t -test should only be undertaken when the variances of X and Y are almost certainly equal.

Example 9.11 ▷ *t-Test, $\sigma_X = \sigma_Y$ Assumed: School Satisfaction* ◁ A questionnaire is devised by the Board of Governors to measure the level of satisfaction for graduates from two competing state schools. Past history indicates the variance in satisfaction levels for both schools is equal. The questionnaire is randomly administered to 11 students from State School X and 15 students from State School Y (the results have been ordered and stored in data frame `Stschool`).

School X: 69 75 76 80 81 82 86 89 91 92 97

School Y: 59 62 66 70 70 75 75 77 78 79 81 84 84 86 94

- Test to see if there are significant differences between the mean satisfaction levels for graduates of the two competing state schools using a significance level of 5%.
- Find the power for $\mu_1(X, Y) = \mu_X - \mu_Y = 10$ for the test in (a) if it is assumed $\sigma_X = \sigma_Y = 9$.

Solution: The answers are as follows:

(a) To solve this part, start by verifying the reasonableness of the normality assumption. The side-by-side boxplots and normal quantile-quantile plots depicted in Figure 9.8 suggest it is reasonable to assume the satisfaction levels for graduates from both state schools follow normal distributions.

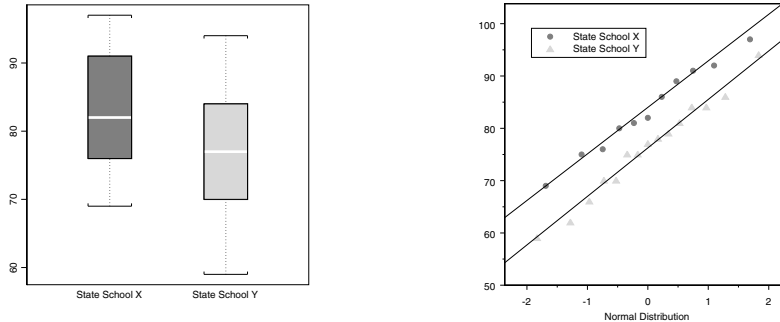


FIGURE 9.8: Side-by-side boxplots and normal quantile-quantile plots of the satisfaction level for graduates from State School X and State School Y .

Five-Step Procedure:

Step 1: **Hypotheses** — Since the problem gives no reason to suspect graduates from School X are any more satisfied than graduates from School Y, use a two-tailed alternative hypothesis:

$$H_0 : \mu_X - \mu_Y = 0 \text{ versus } H_1 : \mu_X - \mu_Y \neq 0$$

Step 2: **Test Statistic** — The test statistic chosen is $\bar{X} - \bar{Y}$ because $E[\bar{X} - \bar{Y}] = \mu_X - \mu_Y$. The value of this test statistic is $83.45 - 76 = 7.45$. The standardized test statistic under the assumption that H_0 is true and its distribution are

$$\frac{[(\bar{X} - \bar{Y}) - \delta_0]}{\sqrt{S_p^2 \left(\frac{1}{n_X} + \frac{1}{n_Y} \right)}} \sim t_{n_X + n_Y - 2}.$$

Step 3: **Rejection Region Calculations** — Because the standardized test statistic is distributed $t_{n_X + n_Y - 2}$, and H_1 is a two-sided hypothesis, the rejection region is $|t_{\text{obs}}| > t_{1-0.05/2; 11+15-2} = t_{0.975; 24} = 2.06$. Note that $s_X = 8.41$, $s_Y = 9.45$, and the pooled standard deviation is $s_p = 9.03$. The value of the standardized test statistic is

$$t_{\text{obs}} = \frac{\bar{x} - \bar{y} - \delta_0}{s_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} = \frac{83.45 - 76 - 0}{9.03 \sqrt{\frac{1}{11} + \frac{1}{15}}} = 2.08.$$

Step 4: **Statistical Conclusion** — The φ -value is $2 \times \mathbb{P}(t_{24} \geq 2.08) = 2 \times 0.024 = 0.048$.

- I. From the rejection region, reject H_0 because $|t_{\text{obs}}| = 2.08$ is greater than 2.06.
- II. From the φ -value, reject H_0 because the φ -value = 0.048 is less than 0.05.

Reject H_0 .

Step 5: **English Conclusion** — There is evidence to suggest the average satisfaction levels between State School X and State School Y are different.

To perform the calculations with S, attach `Stschool` and enter

```
> round(qt(0.975,24), 2)                # Critical Value
[1] 2.06
> mX <- mean(X, na.rm=TRUE)
> mY <- mean(Y, na.rm=TRUE)
> sX <- sd(X, na.rm=TRUE)                # stdev(X, na.rm=T) for S-PLUS
> sY <- sd(Y, na.rm=TRUE)                # stdev(Y, na.rm=T) for S-PLUS
> sp <- sqrt((10*sX^2 + 14*sY^2)/24)      # Pooled stdev
> tobs <- (mX - mY)/(sp*sqrt(1/11 + 1/15)) # t obs
> round(c(mX, mY, sX, sY, sp, tobs), 2)
[1] 83.45 76.00 8.41 9.45 9.03 2.08
> 2*(1 - pt(tobs,24))                    # P-value
[1] 0.04839673
```

To compute the value of the standardized test statistic and its corresponding p -value with S, key in

```
> t.test(X, Y, var.equal=TRUE)
```

Standard Two-Sample t-Test

```
data: X and Y
t = 2.0798, df = 24, p-value = 0.0484
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.05691592 14.85217499
sample estimates:
mean of x mean of y
 83.45455      76
```

The confidence interval is calculated by S using (8.15) on page 308 and does not include 0. Thus, a conclusion based on this interval would be identical to that in step 5 of the five-step procedure used to solve this problem.

(b) Before computing $Power(\mu_1(X, Y) = 10)$, first determine the non-centrality parameter:

$$\gamma = \frac{\mu_1(X, Y) - \mu_0(X, Y)}{\sigma_{\bar{X} - \bar{Y}}} = \frac{10 - 0}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} = \frac{10}{\sqrt{\frac{9^2}{11} + \frac{9^2}{15}}} = \frac{10}{3.573} = 2.80.$$

Let $T = t(\mathbf{X}, \mathbf{Y}) = \frac{\bar{X} - \bar{Y} - \mu_1(X, Y)}{S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}$. Then

$$\begin{aligned}
\text{Power}(\mu_1(X, Y) = 10) &= \mathbb{P}(\text{Reject } H_0 | H_1) \\
&= \mathbb{P}\left((T < t_{\alpha/2; n_X + n_Y - 2}) \mid T \sim t_{n_X + n_Y - 2; \gamma}^*\right) + \\
&\quad \mathbb{P}\left((T > t_{1-\alpha/2; n_X + n_Y - 2}) \mid T \sim t_{n_X + n_Y - 2; \gamma}^*\right) \\
&= \mathbb{P}\left((t_{24; 2.8}^* < t_{0.025; 24})\right) + \mathbb{P}\left((t_{24; 2.8}^* > t_{0.975; 24})\right) \\
&= \mathbb{P}\left((t_{24; 2.8}^* < -2.06)\right) + \mathbb{P}\left((t_{24; 2.8}^* > 2.06)\right) = 0.766
\end{aligned}$$

Find the $\text{Power}(\mu_1(X, Y) = 10)$ using the non-central t -distribution and the non-central F distribution in R. Note that if one is using S-PLUS, one can only solve the problem using the non-central F distribution. To calculate the quantity $\mathbb{P}\left((t_{24; 3}^* < t_{0.025; 24})\right) + \mathbb{P}\left((t_{24; 3}^* > t_{0.975; 24})\right)$ with R, enter

```
> pt(qt(0.025, 24), 24, 2.8) + (1-pt(qt(0.975, 24), 24, 2.8))
[1] 0.7662468
```

Using the relationship between t -distributions and F distributions given in (9.4), write

$$\begin{aligned}
\mathbb{P}\left((t_{24; 2.8}^* < t_{0.025; 24}) \cup (t_{24; 2.8}^* > t_{0.975; 24})\right) &= \mathbb{P}\left(F_{1, 24; \gamma=2.8^2} > (t_{1-\alpha/2; n-1})^2\right) \\
&= \mathbb{P}\left(F_{1, 24; \gamma=2.8^2} > (t_{0.975; 24})^2\right) \\
&= \mathbb{P}\left(F_{1, 24; \gamma=2.8^2} > (2.06)^2 = 4.26\right) \\
&= 0.766.
\end{aligned}$$

To find $\mathbb{P}(F_{1, 24; 9} > 4.26) = 1 - \mathbb{P}(F_{1, 24; 7.84} < 4.26)$ with S, key in

```
> 1-pf(qt(.975, 24)^2, 1, 24, 7.84)
[1] 0.7662468
```

9.7.5 Test for a Difference in Means when Sampling from Independent Normal Distributions with Variances That Are Unknown and Unequal

Recall that when random samples of size n_X and n_Y , respectively, are taken from two normal distributions $N(\mu_X, \sigma)$, and $N(\mu_Y, \sigma)$, where σ is known, the random variable

$$Z = \frac{[(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)]}{\sqrt{\left(\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}\right)}} \sim N(0, 1).$$

In real problems, the values of the population variances are seldom known. Further, the random variable

$$\frac{[(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)]}{\sqrt{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)}}, \quad (9.5)$$

does not have a known distribution. However, the random variable in (9.5) can be approximated with a t -distribution with ν degrees of freedom, where

$$\nu = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\frac{(s_X^2/n_X)^2}{n_X - 1} + \frac{(s_Y^2/n_Y)^2}{n_Y - 1}}. \quad (9.6)$$

The approximation of the random variable (9.5) with a t_ν is known as the **Welch-Satterthwaite** method. Output from `S` using this technique is simply labeled **Welch**.

The null hypothesis used to test for a difference of means between two independent normal distributions where the variances are unknown and unequal is $H_0 : \mu_X - \mu_Y = \delta_0$. The value of the standardized test statistic using the Welch-Satterthwaite method is written

$$t_{\text{obs}} = \frac{\bar{x} - \bar{y} - \delta_0}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}}.$$

The three possible alternative hypothesis and the corresponding rejection regions are in Table 9.10.

Table 9.10: Summary for test for differences in mean when taking independent samples from normal distributions with unknown and unequal variances (Welch test)

Null Hypothesis — $H_0 : \mu_X - \mu_Y = \delta_0$

Standardized Test
Statistic's Value — $t_{\text{obs}} = \frac{\bar{x} - \bar{y} - \delta_0}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}}$

Alternative Hypothesis	Rejection Region
$H_1 : \mu_X - \mu_Y < \delta_0$	$t_{\text{obs}} < t_{\alpha; \nu}$
$H_1 : \mu_X - \mu_Y > \delta_0$	$t_{\text{obs}} > t_{1-\alpha; \nu}$
$H_1 : \mu_X - \mu_Y \neq \delta_0$	$ t_{\text{obs}} > t_{1-\alpha/2; \nu}$

Example 9.12 A bottled water company acquires its water from two independent sources, X and Y. The company suspects that the sodium content in the water from source X is less than the sodium content for water from source Y. An independent agency measures the sodium content in 20 samples from source X and 10 samples from source Y and stores them in data frame `Water`. Is there statistical evidence to suggest the average sodium content in the water from source X is less than the average sodium content in the water from source Y? The measurements for the sodium values are mg/L. Use an α level of 0.05 to test the appropriate hypotheses.

Source X: 84 73 92 84 95 74 80 86 80 77
 86 72 62 54 77 63 85 59 66 79
 Source Y: 78 79 84 82 80 85 81 83 79 81

Solution: To solve this problem, start by verifying the reasonableness of the normality assumption. The side-by-side boxplots and normal quantile-quantile plots depicted in Figure 9.9 on the facing page suggest it is reasonable to assume the sodium values for both sources follow normal distributions; however, it is clear from the boxplot that the variances are very different. Now, proceed with the five-step procedure.

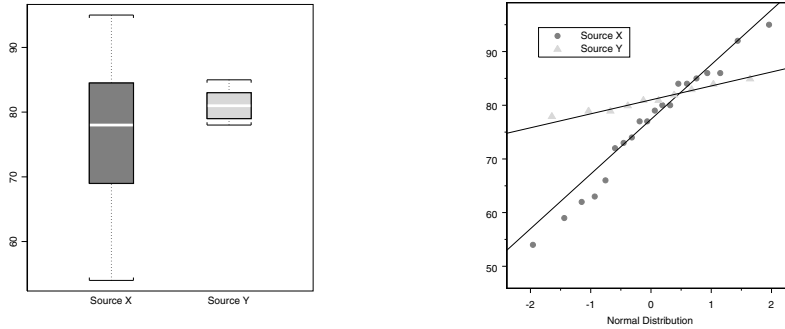


FIGURE 9.9: Side-by-side boxplots and normal quantile-quantile plots of the sodium content for source X and source Y.

Step 1: **Hypotheses** — Since the problem wants to test to see if the mean sodium content from source X is less than the mean sodium content from source Y, use a lower one-sided alternative hypothesis.

$$H_0 : \mu_X - \mu_Y = 0 \text{ versus } H_1 : \mu_X - \mu_Y < 0$$

Step 2: **Test Statistic** — The test statistic chosen is $\bar{X} - \bar{Y}$ because $E[\bar{X} - \bar{Y}] = \mu_X - \mu_Y$. The value of this test statistic is $76.4 - 81.2 = -4.8$. The standardized test statistic under the assumption that H_0 is true and its approximate distribution are

$$\frac{[(\bar{X} - \bar{Y}) - \delta_0]}{\sqrt{\left(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}\right)}} \sim t_\nu.$$

Step 3: **Rejection Region Calculations** — Because the standardized test statistic is distributed approximately t_ν , and H_1 is a lower one-sided hypothesis, the rejection region is $t_{\text{obs}} < t_{0.05;22} = -1.72$. The degrees of freedom are

$$\nu = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\frac{(s_X^2/n_X)^2}{n_X-1} + \frac{(s_Y^2/n_Y)^2}{n_Y-1}} = \frac{\left(\frac{122.78}{20} + \frac{5.29}{10}\right)^2}{\frac{(122.78/20)^2}{20-1} + \frac{(5.29/10)^2}{10-1}} = 22.07,$$

and the value of the standardized test statistic is

$$t_{\text{obs}} = \frac{\bar{x} - \bar{y} - \delta_0}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}} = \frac{76.4 - 81.2 - 0}{\sqrt{\frac{122.78}{20} + \frac{5.29}{10}}} = -1.86.$$

Step 4: **Statistical Conclusion** — The φ -value is $\mathbb{P}(t_{22} \leq -1.86) = 0.038$.

- I. From the rejection region, reject H_0 because $t_{\text{obs}} = -1.86$ is less than -1.72 .
- II. From the φ -value, reject H_0 because the φ -value = 0.038 is less than 0.05.

Reject H_0 .

Step 5: **English Conclusion** — There is evidence to suggest the average sodium content for source X is less than the average sodium content for source Y.

To perform the calculations with S, attach `Water` and enter

```
> X <- X[!is.na(X)]
> Y <- Y[!is.na(Y)]
> nX <- length(X); nY <- length(Y); mX <- mean(X)
> mY <- mean(Y); sX2 <- var(X); sY2 <- var(Y)
> nu <- (sX2/nX + sY2/nY)^2/((sX2/nX)^2/(nX-1) + (sY2/nY)^2/(nY-1))
> round(qt(0.05, nu), 2) # Critical Value
[1] -1.72
> tobs <- (mX - mY)/sqrt(sX2/nX + sY2/nY) # t observed
> round(c(mX, mY, sX2, sY2, nu, tobs), 2)
[1] 76.40 81.20 122.78 5.29 22.07 -1.86
> pt(tobs, nu) # P-value
[1] 0.03821647
```

To compute the value of the standardized test statistic and its corresponding ϕ -value with S, type

```
> t.test(X, Y, var.equal=FALSE, alternative="less")
```

Welch Modified Two-Sample t-Test

```
data: X and Y
t = -1.8589, df = 22.069, p-value = 0.0382
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      NA -0.3665724
sample estimates:
 mean of x mean of y
      76.4      81.2
```

The confidence interval S calculates agrees with the one from (8.16) on page 310 modified for a one-sided confidence interval. Note that the values included in the confidence interval are all less than zero, which would give a conclusion identical to that found in step 5 of the five-step procedure. ■

9.7.6 Test for the Mean Difference when the Differences Have a Normal Distribution

If one wants to test whether there has been some change in a single group of subjects or if there exists some difference between two dependent samples, one can compute the net change from one condition to the next and do a paired t -test provided certain normality assumptions are satisfied. Recall from Section 8.2.7 on page 313 that when a researcher is presented with paired samples, the standard approach is to analyze the differences between the paired data. Provided the distribution of population differences is

$$D \sim N(\mu_D = \mu_X - \mu_Y, \sigma_D),$$

the random variable

$$T = \frac{\bar{D} - \mu_D}{S_D/\sqrt{n_D}} \sim t_{n-1}.$$

The null hypothesis for testing a difference of means with dependent samples is $H_0 : \mu_D = \mu_X - \mu_Y = \delta_0$, and the value of the standardized test statistic is written

$$t_{\text{obs}} = \frac{\bar{d} - \delta_0}{s_D / \sqrt{n_D}}.$$

The three alternative hypotheses and the rejection regions for H_0 are in Table 9.11. The paired t -test has a smaller variance than does an independent two-sample t -test when the data are dependent and is a special case of the experimental design known as the randomized block design. The matched differences are known as **blocks**. Blocks should be used any time the differences within a block are relatively homogeneous compared to the differences within the particular treatment. When blocks are used appropriately, differences noted in the paired observations can subsequently be attributed to differences in treatments.

Table 9.11: Summary for testing the mean of the differences between two dependent samples when the differences follow a normal distribution with unknown variance (paired t -test)

Null Hypothesis — $H_0 : \mu_D = \mu_X - \mu_Y = \delta_0$

Standardized
Test Statistic's
Value — $t_{\text{obs}} = \frac{\bar{d} - \delta_0}{s_D / \sqrt{n_D}}$

Alternative Hypothesis	Rejection Region
$H_1 : \mu_D < \delta_0$	$t_{\text{obs}} < t_{\alpha; n-1}$
$H_1 : \mu_D > \delta_0$	$t_{\text{obs}} > t_{1-\alpha; n-1}$
$H_1 : \mu_D \neq \delta_0$	$ t_{\text{obs}} > t_{1-\alpha/2; n-1}$

Example 9.13 The data frame `barley` in `S-PLUS` or in the `lattice` package lists barley yield in bushels per acre for the years 1931 and 1932 for ten varieties of barley grown at six sites. Is there evidence to suggest the average barley yield in 1932 for the Morris site is greater than the average barley yield in 1932 for the Crookston site? Use the five-step procedure to test the appropriate hypotheses using an $\alpha = 0.05$ significance level.

Solution: Note that the same ten varieties are grown at both the Morris and the Crookston site. Consequently, the yields at the two sites are dependent on the varieties. That is, variety acts as a block. It stands to reason that one can expect less variability between two similar plots growing the same variety than the variability within each of the plots growing different varieties. Start the analysis by verifying the normality assumption required to use a paired t -test. The results from applying the function `EDA()` to the differences between the 1932 barley yields from the Morris and Crookston sites are provided in Figure 9.10 on the following page. Based on the graphical output from the function `EDA()`, it is not unreasonable to assume the differences between the 1932 barley yields from the Morris and Crookston sites follow a normal distribution. Now, proceed with the five-step procedure.

Step 1: **Hypotheses** — To test if the average 1932 barley yield from Morris is greater than the average 1932 barley yield from Crookston, the hypotheses are

$$H_0 : \mu_D = 0 \text{ versus } H_1 : \mu_D > 0$$

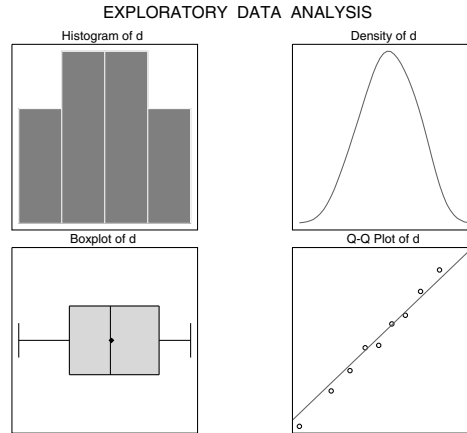


FIGURE 9.10: Exploratory data analysis of the differences between 1932 barley yields from the Morris and Crookston sites.

Step 2: **Test Statistic** — The test statistic chosen is \bar{D} because $E[\bar{D}] = \mu_D$. The value of this test statistic is $\bar{d} = 10.33$. The standardized test statistic under the assumption that H_0 is true and its distribution are $\frac{\bar{D} - \delta_0}{s_D / \sqrt{n_D}} \sim t_{10-1}$.

Step 3: **Rejection Region Calculations** — Because the standardized test statistic is distributed t_9 , and H_1 is an upper one-sided hypothesis, the rejection region is $t_{\text{obs}} > t_{1-0.05;9} = t_{0.95;9} = 1.83$. The value of the standardized test statistic is $t_{\text{obs}} = \frac{\bar{d} - \delta_0}{s_D / \sqrt{n_D}} = \frac{10.33 - 0}{5.19 / \sqrt{10}} = 6.29$.

Step 4: **Statistical Conclusion** — The φ -value is $\mathbb{P}(t_9 \geq 6.29) \approx 0$.

- I. From the rejection region, reject H_0 because $t_{\text{obs}} = 6.29$ is greater than 1.83.
- II. From the φ -value, reject H_0 because the φ -value ≈ 0 is less than 0.05.

Reject H_0 .

Step 5: **English Conclusion** — There is evidence to suggest that the 1932 mean barley yield for Morris is greater than the 1932 mean barley yield for Crookston.

To compute the rejection region, value of the standardized test statistic, and its corresponding φ -value with S, enter

```
> qt(.95,9)                # Critical Value
[1] 1.833113
> library(lattice)         # Not needed for S-PLUS
> attach(barley)
> yieldMor32 <- yield[year == "1932" & site == "Morris"]
> yieldCro32 <- yield[year == "1932" & site == "Crookston"]
> d <- yieldMor32 - yieldCro32
```

```
> t.test(d, alternative = "greater")
```

```
One-sample t-Test
```

```
data: d
t = 6.2924, df = 9, p-value = 0.0001
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 7.323012      NA
sample estimates:
mean of x
10.33333
```

An alternative method that yields identical results is to use the command

```
t.test(yieldMor32, yieldCro32, paired=TRUE, alternative="greater").
```

Note that the confidence interval calculated by `S` is using (8.21) on page 314 modified for one-sided confidence intervals. The interval calculated agrees with our conclusion from step 5 because it contains values that are exclusively greater than zero. ■

9.8 Hypothesis Tests for Population Variances

9.8.1 Test for the Population Variance when Sampling from a Normal Distribution

The tests for population means presented up to this point have assumed the sampling distributions for their corresponding statistics follow a normal distribution. However, the tests for means are fairly robust to violations in normality assumptions. In contrast, the normality assumption for testing a hypothesis about variance is not robust to departures from normality. Consequently, one should proceed with caution when testing a hypothesis about the variance especially since non-normality is difficult to detect when working with small to moderate size samples. As a minimum, one should look at a normal quantile-quantile plot to make sure normality is plausible before testing a hypothesis concerning the population variance.

Provided X_1, X_2, \dots, X_n is a random sample from a $N(\mu, \sigma)$ distribution, the random variable

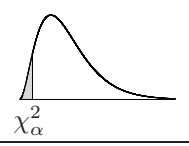
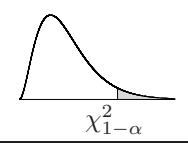
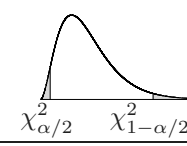
$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2.$$

The null hypothesis for testing the population variance is $H_0 : \sigma^2 = \sigma_0^2$, and the value for the test statistic is

$$\chi_{\text{obs}}^2 = \frac{(n-1)s^2}{\sigma_0^2}.$$

The three alternative hypotheses and the rejection regions for H_0 are in Table 9.12 on the following page.

Table 9.12: Summary for testing the population variance when sampling from a normal distribution

Null Hypothesis — $H_0 : \sigma^2 = \sigma_0^2$	Standardized Test Statistic's Value — $\chi_{\text{obs}}^2 = \frac{(n-1)s^2}{\sigma_0^2}$		
Alternative Hypothesis	$H_1 : \sigma^2 < \sigma_0^2$	$H_1 : \sigma^2 > \sigma_0^2$	$H_1 : \sigma^2 \neq \sigma_0^2$
Rejection Region	$\chi_{\text{obs}}^2 < \chi_{\alpha; n-1}^2$	$\chi_{\text{obs}}^2 > \chi_{1-\alpha; n-1}^2$	$\chi_{\text{obs}}^2 < \chi_{\alpha/2; n-1}^2 \cup \chi_{\text{obs}}^2 > \chi_{1-\alpha/2; n-1}^2$
Graphical Representation of Rejection Region			
Note that the degrees of freedom for all the χ^2 values are $n - 1$.			

Example 9.14 The quality control office of a large hardware manufacturer received more than twice the number of complaints it usually receives in reference to the diameter variability of its 4 cm washers. In light of the complaints, the quality control manager wants to ascertain whether or not there has been an increase in the diameter variability of the company's washers manufactured this month versus last month, where the variance was 0.004 cm^2 . The manager takes a random sample of 20 washers manufactured this month. The results are recorded in Table 9.13 and stored in the data frame `Washer`. Conduct an appropriate hypothesis test using a significance level of $\alpha = 0.05$.

Table 9.13: Diameters for 20 randomly selected washers (`Washer`)

4.06	4.02	4.04	4.04	3.97	3.87	4.03	3.85	3.91	3.98
3.96	3.90	3.95	4.11	4.00	4.12	4.00	3.98	3.92	4.02

Solution: Prior to using a test that is very sensitive to departures in normality, as a minimum, create a quantile-quantile plot to verify the assumption of normality. The results from applying `EDA()`, shown in Figure 9.11 on the facing page, suggest the diameters of the washers follow a normal distribution. Now, continue with the five-step procedure.

Step 1: Hypotheses — The null and alternative hypotheses to test whether the diameter variability of the companies washers manufactured this month is greater than the variability last month, where the variance was 0.004 cm^2 , are

$$H_0 : \sigma^2 = 0.004 \text{ versus } H_1 : \sigma^2 > 0.004.$$

Step 2: Test Statistic — The test statistic chosen is S^2 because $E[S^2] = \sigma^2$. The value of this test statistic is $s^2 = 0.005318684$. The standardized test statistic under the assumption that H_0 is true and its distribution are $\frac{(n-1)S^2}{\sigma_0^2} \sim \chi_{n-1}^2$.

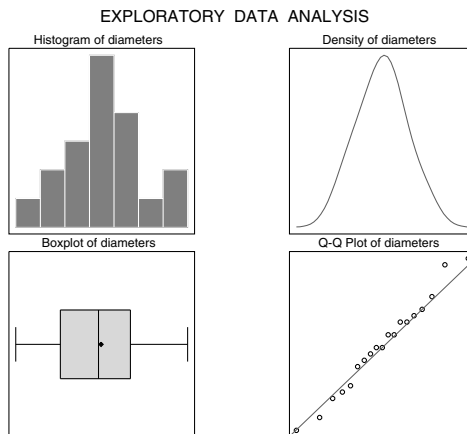


FIGURE 9.11: Graphs from using `EDA()` on the washers' diameters

Step 3: **Rejection Region Calculations** — Because the standardized test statistic is distributed χ_{19}^2 , and H_1 is an upper one-sided hypothesis, the rejection region is $\chi_{\text{obs}}^2 > \chi_{0.95;19}^2 = 30.14$. The value of the standardized test statistic is $\chi_{\text{obs}}^2 = \frac{(20-1)(0.005318684)}{0.004} = 25.26$.

Step 4: **Statistical Conclusion** — The φ -value is $\mathbb{P}(\chi_{19}^2 \geq 25.27) = 0.15$.

- I. From the rejection region, fail to reject H_0 because $\chi_{\text{obs}}^2 = 25.27$ is less than 30.14.
- II. From the φ -value, fail to reject H_0 because the φ -value = 0.15 is greater than 0.05.

Fail to reject H_0 .

Step 5: **English Conclusion** — There is insufficient evidence to suggest the variance for washers manufactured this month increased from the variance of washers manufactured last month.

To compute the critical value, the standardized test statistic's value, and the corresponding φ -value with S, use the variable `washers`, which contains the variance of last month's washers' diameters.

```
> qchisq(0.95,19)           # Critical Value
[1] 30.14351
> attach(Washer)
> s2 <- var(diameters)
> s2
[1] 0.005318684
> ChiObs <- 19*s2/0.004     # Standardized Test Statistic's Value
> ChiObs
[1] 25.26375
> 1-pchisq(ChiObs,19)     # P-value
[1] 0.1520425
```



9.8.2 Test for Equality of Variances when Sampling from Independent Normal Distributions

This section addresses the issue of comparing the variances of two distributions. This problem is encountered when comparing instrument precisions or uniformity of products. Another application is to check the assumption of equal variances for the pooled t -test. However, as mentioned earlier, the pooled t -test should only be used when equality of variance is beyond doubt. Consequently, this text will not place as large an emphasis on this use of the test as some other texts do. Provided X_1, X_2, \dots, X_{n_X} and Y_1, Y_2, \dots, Y_{n_Y} are independent random samples from $N(\mu_X, \sigma_X)$ and $N(\mu_Y, \sigma_Y)$ distributions, respectively, the random variable

$$F = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F_{n_X-1, n_Y-1}.$$

The null hypothesis for testing the equality of two population variances is $H_0 : \sigma_X^2 = \sigma_Y^2$, which is equivalent to testing $H_0 : \sigma_X^2/\sigma_Y^2 = 1$. The value for the test statistic when the variances are assumed equal is

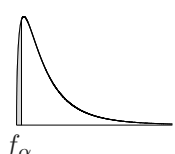
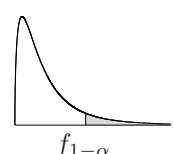
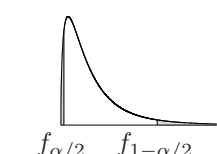
$$f_{\text{obs}} = \frac{s_X^2}{s_Y^2}.$$

The three alternative hypotheses and the rejection regions for H_0 are in Table 9.14.

Table 9.14: Summary for test for equality of variances when sampling from independent normal distributions

Null Hypothesis — $H_0 : \sigma_X^2 = \sigma_Y^2$ Standardized Test Statistic's Value — $f_{\text{obs}} = \frac{s_X^2}{s_Y^2}$

Note that all f values in this table have degrees of freedom $n_X - 1, n_Y - 1$.

Alternative Hypothesis	$H_1 : \sigma_X^2 < \sigma_Y^2$	$H_1 : \sigma_X^2 > \sigma_Y^2$	$H_1 : \sigma_X^2 \neq \sigma_Y^2$
Rejection Region	$f_{\text{obs}} < f_\alpha$	$f_{\text{obs}} > f_{1-\alpha}$	$f_{\text{obs}} < f_{\alpha/2}$ OR $f_{\text{obs}} > f_{1-\alpha/2}$
Graphical Representation of Rejection Region			

Example 9.15 ▷ *F-Test: Breathalyzers* ◁ In an effort to reduce the number of drunk drivers associated with fraternal organizations, the fraternity council wants to distribute portable breathalyzers to all the fraternities on campus. There are two companies that are bidding to provide these breathalyzers. The fraternity council has decided to purchase all of its breathalyzers from the company whose breathalyzers have the smaller variance. Based on advertisement, the fraternity council suspects breathalyzer X to have a smaller variance than breathalyzers from company Y. Each company provides ten portable breathalyzers to the fraternity council. Two volunteers each consumed a 12-ounce beer every 15 minutes for one hour. One hour after the fourth beer was consumed, each volunteer's blood alcohol was measured with a different breathalyzer from the same company. The numbers recorded in

data frame `Bac` are the sorted blood alcohol content values reported with breathalyzers from company X and company Y. Test the appropriate hypotheses using a 5% significance level. (Note: The units of measurement for blood alcohol content, BAC, are grams of alcohol per liter of blood, g/L .)

Company X: 0.08 0.09 0.09 0.10 0.10 0.10 0.10 0.11 0.11 0.12

Company Y: 0.00 0.03 0.04 0.04 0.05 0.05 0.06 0.07 0.08 0.08

Solution: Prior to using a test that is very sensitive to departures in normality, the function `EDA()` is applied to the ten blood alcohol readings using breathalyzers from company X and the ten blood alcohol readings recorded using breathalyzers from company Y. Based on the results displayed in Figure 9.12, it seems reasonable to assume the blood alcohol values breathalyzers report from both companies X and Y follow normal distributions. Although the blood alcohol values reported with company Y analyzers are slightly skewed to the left, one must remember that only ten values were used in the construction of the graphs, and that graphs constructed with small numbers even when sampling from normal distributions will often appear skewed. When working with small sample sizes, one may want to test formally the hypothesis of normality with a function like `shapiro.test()`, which is explained more fully in Section 10.7.3 of Chapter 10. The Shapiro-Wilk Normality Test also indicates normality is plausible based on the relatively large p -value (0.5489). Therefore, proceed with the five-step procedure.

```
> attach(Bac)
> shapiro.test(Y)
```

Shapiro-Wilk Normality Test

```
data: Y
W = 0.9396, p-value = 0.5489
```

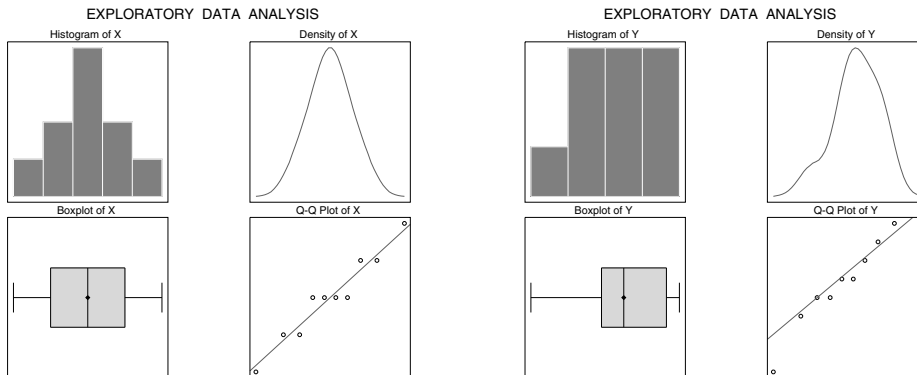


FIGURE 9.12: Exploratory data analysis for the blood alcohol values using the breathalyzers from company X and company Y on two volunteers after drinking four beers.

Step 1: **Hypotheses** — The null and alternative hypotheses to test whether the variability in blood alcohol values using company X's breathalyzers is less than the variability

in blood alcohol values using company Y's breathalyzers are

$$H_0 : \sigma_X^2 = \sigma_Y^2 \text{ versus } H_1 : \sigma_X^2 < \sigma_Y^2.$$

Step 2: **Test Statistic** — The test statistics chosen are S_X^2 and S_Y^2 since $E[S_X^2] = \sigma_X^2$ and $E[S_Y^2] = \sigma_Y^2$. The values of these test statistics are $s_X^2 = 0.0001333333$ and $s_Y^2 = 0.0006$. The standardized test statistic under the assumption that H_0 is true and its distribution are $S_X^2/S_Y^2 \sim F_{10-1,10-1}$.

Step 3: **Rejection Region Calculations** — Because the standardized test statistic is distributed $F_{9,9}$, and H_1 is a lower one-sided hypothesis, the rejection region is $f_{\text{obs}} < F_{0.05;9,9} = 0.31$. The value of the standardized test statistic is $f_{\text{obs}} = (0.0001333333)/(0.0006) = 0.2222$.

Step 4: **Statistical Conclusion** — The φ -value is $\mathbb{P}(F_{9,9} \leq 0.2222) = 0.0176$.

I. From the rejection region, reject H_0 because $f_{\text{obs}} = 0.2222$ is less than 0.31.

II. From the φ -value, reject H_0 because the φ -value = 0.0176 is less than 0.05.

Reject H_0 .

Step 5: **English Conclusion** — The evidence suggests the variability of blood alcohol values using breathalyzers from company X is less than the variance for blood alcohol values using breathalyzers from company Y.

To compute the critical value, the standardized test statistic's value, and the corresponding φ -value with S, enter

```
> qf(.05,9,9)           # Critical Value
[1] 0.3145749
> var(X)
[1] 0.0001333333
> var(Y)
[1] 0.0006
> var(X)/var(Y)
[1] 0.2222222           # Standardized Test Statistic's Value
> pf(var(X)/var(Y),9,9) # P-value
[1] 0.01764349
```

To test the appropriate hypothesis using the S function `var.test()`, key in

```
> var.test(X, Y, alternative="less")
```

F test to compare two variances

data: X and Y

F = 0.2222, num df = 9, denom df = 9, p-value = 0.01764

alternative hypothesis: true ratio of variances is less than 1

95 percent confidence interval:

0.0000000 0.7064207

sample estimates:

ratio of variances

0.2222222

The confidence interval is calculated with (8.31) on page 319, modified for a one-sided confidence interval. Note that the interval agrees with our step 5 conclusion as it contains values that are exclusively less than 1, implying $\frac{\sigma_x^2}{\sigma_y^2} < 1$. ■

9.9 Hypothesis Tests for Population Proportions

9.9.1 Testing the Proportion of Successes in a Binomial Experiment (Exact Test)

Tests of hypotheses concerning proportions are encountered in many areas. Two examples are manufacturing firms often test the percent of defective items in their products and politicians test that their support base will garner them a certain proportion of votes in an election. Many other examples exist. In this section, the problem of testing a hypothesis where the proportion of successes in a binomial experiment (π) is equal to some value (π_0) is considered. Specifically, given a random variable $Y \sim \text{Bin}(n, \pi)$, an exact test for the null hypothesis $H_0 : \pi = \pi_0$, is constructed. The three possible alternative hypotheses and the \wp -value formulas associated with each alternative hypothesis are given in Table 9.15. Note the use of the indicator function in the computation of the two-sided \wp -value. Although it is possible to calculate rejection regions for this exact test, due to the discrete nature of Y , it is unlikely a critical region can be established whose size is exactly equal to the prescribed α . Therefore, \wp -values are generally computed when working with exact problems instead of defining rejection regions.

Table 9.15: Summary for testing the proportion of successes in a binomial experiment (number of successes is $Y \sim \text{Bin}(n, \pi)$)

Null Hypothesis — $H_0 : \pi = \pi_0$

Test Statistic's Value — $y_{\text{obs}} = \text{number of observed successes}$

Alternative Hypothesis	\wp -Value Formula
$H_1 : \pi < \pi_0$	$\mathbb{P}(Y \leq y_{\text{obs}} H_0) = \sum_{i=0}^{y_{\text{obs}}} \binom{n}{i} \pi_0^i (1 - \pi_0)^{n-i}$
$H_1 : \pi > \pi_0$	$\mathbb{P}(Y \geq y_{\text{obs}} H_0) = \sum_{i=y_{\text{obs}}}^n \binom{n}{i} \pi_0^i (1 - \pi_0)^{n-i}$
$H_1 : \pi \neq \pi_0$	$\sum_{i=0}^n I(\mathbb{P}(Y = i) \leq \mathbb{P}(Y = y_{\text{obs}})) \cdot \binom{n}{i} \pi_0^i (1 - \pi_0)^{n-i}$

It is also possible to compute an exact confidence interval for π . However, due to the discrete nature of Y , the actual confidence level (coverage probability) of the interval is often considerably higher than the stated confidence level. An exact $(1 - \alpha) \cdot 100\%$ confidence interval for π requires each one-sided \wp -value in an exact binomial test to exceed $\alpha/2$. Except when $y = 0$ and the lower bound is zero, and when $y = n$ and the upper bound is

1, the lower and upper endpoints for an exact $(1 - \alpha) \cdot 100\%$ confidence interval for π are the solutions in π_0 to the equations

$$\sum_{k=0}^y \binom{n}{k} \pi_0^k (1 - \pi_0)^{n-k} \geq \frac{\alpha}{2} \quad \text{and} \quad \sum_{k=y}^n \binom{n}{k} \pi_0^k (1 - \pi_0)^{n-k} \geq \frac{\alpha}{2}. \quad (9.7)$$

For values of $y = 1, 2, \dots, n - 1$, it can be shown that the solutions to (9.7) yield the lower and upper endpoint expressions for the confidence interval given in

$$CI_{1-\alpha}(\pi) = \left[\left(1 + \frac{n - y_{\text{obs}} + 1}{y_{\text{obs}} F_{\alpha/2; 2y_{\text{obs}}, 2(n-y_{\text{obs}}+1)}} \right)^{-1}, \left(1 + \frac{n - y_{\text{obs}}}{(y_{\text{obs}} + 1) F_{1-\alpha/2; 2(y_{\text{obs}}+1), 2(n-y_{\text{obs}})}} \right)^{-1} \right] \quad (9.8)$$

Both **S-PLUS** and **R** perform an exact binomial test using the function `binom.test()`. However, at the time of writing, the **S-PLUS** `binom.test()` did not provide a corresponding confidence interval and uses a different criterion to compute its φ -value for two-sided alternatives than the one presented in Table 9.15 on the previous page. The **R** `binom.test()` uses the criterion in Table 9.15 on the preceding page, called the **likelihood** method, to compute its φ -values for two-sided alternatives, while the **S-PLUS** `binom.test()` uses a **tail-balancing** criterion; see Blaker (2000). Using the tail-balancing approach, the φ -value is the minimum of the two-tailed probabilities $\mathbb{P}(Y \geq y_{\text{obs}})$ and $\mathbb{P}(Y \leq y_{\text{obs}})$ plus an attainable probability in the other tail that is as close as possible to, but not greater than, that one-tailed probability. The criteria both **S-PLUS** and **R** use to compute φ -values with `binom.test()` for two-sided alternatives differ from the general criterion of

$$2 \min[\mathbb{P}(Y \leq y_{\text{obs}} | H_0), \mathbb{P}(Y \geq y_{\text{obs}} | H_0)],$$

used up to now with two-sided alternatives and continuous distributions. It is of interest to note that the general criterion can be used with the `binom.test()` for two-sided tests and that the φ -values for the likelihood method, the tail-balancing method, and the criterion in Table 9.15 on the previous page will all agree when the distribution is symmetric, that is, when $\pi = 0.5$.

Example 9.16 ▷ *Exact Binomial Test: Graduates' Jobs* ◁ A recent report claimed that 20% of all college graduates find a job in their chosen field of study. A random sample of 500 graduates found that 90 obtained work in their field.

- Is there statistical evidence to refute the claim at the $\alpha = 0.05$ level?
- Compute an exact 95% confidence interval for the true proportion of college graduates that find work in their chosen field of study.

Solution: The answers are as follows:

- Use the five-step procedure.

Step 1: **Hypotheses** — The null and alternative hypotheses to test whether or not 20% of college graduates find work in their chosen field are

$$H_0 : \pi = 0.20 \text{ versus } H_1 : \pi \neq 0.20.$$

Step 2: **Test Statistic** — The test statistic chosen is Y , where Y is the number of college graduates finding work in their chosen field. Provided H_0 is true, $Y \sim \text{Bin}(n, \pi_0)$. The value of the test statistic is $y_{\text{obs}} = 90$.

Step 3: **Rejection Region Calculations** — Rejection is based on the φ -value, so none are required.

Step 4: **Statistical Conclusion — Likelihood Method:**

$$\begin{aligned} \varphi\text{-value} &= \sum_{i=0}^n I(\mathbb{P}(Y = i) \leq \mathbb{P}(Y = y_{\text{obs}})) \cdot \binom{n}{i} \pi_0^i (1 - \pi_0)^{n-i} \\ &= \sum_{i=0}^{500} I(\mathbb{P}(Y = i) \leq \mathbb{P}(Y = 90)) \cdot \binom{500}{i} 0.20^i (1 - 0.20)^{500-i} \\ &= 0.2880566 \quad \text{Computed with S} \end{aligned}$$

Thus, one fails to reject H_0 because 0.2880566 is greater than 0.05.

The S code to compute this φ -value and the R output from using `binom.test()` are

```
> probs <- dbinom(0:500, 500, .2)
> pvalue <- sum(probs[probs <= dbinom(90, 500, .2)])
> pvalue
[1] 0.2880566
> binom.test(x=90, n=500, p=0.2)    # R
```

Exact binomial test

```
data: 90 and 500
number of successes = 90, number of trials = 500, p-value = 0.2881
```

```
alternative hypothesis: true probability of success is not equal to 0.2
```

```
95 percent confidence interval:
```

```
0.1473006 0.2165364
```

```
sample estimates:
```

```
probability of success
```

```
0.18
```


Tail-Balancing Method: To compute the ϕ -value, first find $\mathbb{P}(Y \leq y_{\text{obs}}|H_0)$ and $\mathbb{P}(Y \geq y_{\text{obs}}|H_0)$:

$$\begin{aligned}\mathbb{P}(Y \leq y_{\text{obs}}|H_0) &= \sum_{i=0}^{y_{\text{obs}}} \binom{n}{i} \pi_0^i (1 - \pi_0)^{n-i} \\ &= \sum_{i=0}^{90} \binom{500}{i} 0.20^i (1 - 0.20)^{500-i} \\ &= 0.1437028 \\ \mathbb{P}(Y \geq y_{\text{obs}}|H_0) &= \sum_{i=y_{\text{obs}}}^n \binom{n}{i} \pi_0^i (1 - \pi_0)^{n-i} \\ &= \sum_{i=90}^{500} \binom{500}{i} 0.20^i (1 - 0.20)^{500-i} \\ &= 0.8807233\end{aligned}$$

Since the ϕ -value is computed as $\min\{0.1437028, 0.8807233\} +$ “attainable probability in the other tail,” it follows that the ϕ -value is $0.1437028 + 0.1209751 = 0.2646779$. Thus, one fails to reject H_0 because 0.2646779 is greater than 0.05 .

The S code to compute these ϕ -values is as follows:

```
> p <- pbinom(0:500,500,.2)
> pl <- pbinom(90,500,.2)
> pr <- (1-min(p[p > 1- pl]))
> pval <- pl + pr
> ps <- c(pl, pr, pval)
> ps
[1] 0.1437028 0.1209751 0.2646779
```

The S-PLUS output from using `binom.test()` is

```
> binom.test(90,500,.2)      #S-PLUS

Exact binomial test

data:  90 out of 500
number of successes = 90, n = 500, p-value = 0.2647
alternative hypothesis: true p is not equal to 0.2
```

Fail to reject H_0 .

Step 5: **English Conclusion** — There is not sufficient evidence to suggest the proportion of college graduates finding work in their chosen fields of study is something other than 20%.

(b) An exact 95% confidence interval is constructed using (9.8):

$$\begin{aligned} CI_{1-0.05}(\pi) &= \left[\left(1 + \frac{500 - 90 + 1}{(90)F_{0.05/2;2(90),2(500-90+1)}} \right)^{-1}, \right. \\ &\quad \left. \left(1 + \frac{500 - 90}{(90 + 1)F_{1-0.05/2;2(90+1),2(500-90)}} \right)^{-1} \right] \\ &= \left[\left(1 + \frac{411}{(90)(0.7888743)} \right)^{-1}, \left(1 + \frac{410}{(91)(1.245244)} \right)^{-1} \right] \\ &= [0.1473006, 0.2165364] \end{aligned}$$

One is 95% confident that the true proportion of college graduates finding work in their chosen fields of study lies in $[0.147, 0.216]$. Note that this confidence interval, calculated in step 4 by R, contains the hypothesized value of 0.20, corroborating the decision to fail to reject the null hypothesis. ■

9.9.2 Testing the Proportion of Successes in a Binomial Experiment (Normal Approximation)

In Section 9.9.1, an exact test and confidence interval were presented for the proportion of successes in a binomial experiment where the random variable $Y \sim Bin(n, \pi)$. Specifically, $Y = \sum_{i=1}^n X_i$, where $X_i \sim Bernoulli(\pi)$. A discussion of the properties of Y can be found in Section 6.5.3, starting on page 220. The numerical computations required by exact methods make a computer essentially indispensable, especially when presented with a large sample. Fortunately, for those who do not have access to a computer, approximations to exact distributions are possible for large samples. This is the focus of the current section.

Recall that the asymptotic properties of MLE estimators allow one to write

$$P = \frac{Y}{n} \sim N \left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}} \right) \text{ as } n \rightarrow \infty.$$

Provided $n\pi$ and $n(1-\pi)$ are both greater than or equal to 10,

$$P \dot{\sim} N \left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}} \right) \quad (9.9)$$

provides a reasonable approximation to the sampling distribution of P . Using (9.9), the standardized test statistic under the assumption that $H_0 : \pi = \pi_0$ is true is

$$Z = \frac{P - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \dot{\sim} N(0, 1).$$

The formula to calculate the test statistic's observed value as well as the three possible alternative hypotheses and their rejection regions are described in Table 9.16 on the next page.

When $|p - \pi_0| > \frac{1}{2n}$, many statisticians advocate using a continuity correction when calculating confidence intervals and standardized test statistics' values. A continuity correction of $\pm \frac{1}{2n}$ is automatically applied when using the S function `prop.test()`; however,

Table 9.16: Summary for testing the proportion of successes in a binomial experiment (normal approximation)

Null Hypothesis — $H_0 : \pi = \pi_0$ Standardized Test
Statistic's Value — $z_{\text{obs}} = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}}$

Alternative Hypothesis	Rejection Region
$H_1 : \pi < \pi_0$	$z_{\text{obs}} < z_\alpha$
$H_1 : \pi > \pi_0$	$z_{\text{obs}} > z_{1-\alpha}$
$H_1 : \pi \neq \pi_0$	$ z_{\text{obs}} > z_{1-\alpha/2}$
When $ p - \pi_0 > \frac{1}{2n}$, use a correction factor as in Table 9.17.	

not all statisticians recommend the use of a continuity correction with this test, and using one does lead to a more conservative test. The continuity corrections that are applied, as well as the standardized test statistic calculations, can be found in Table 9.17.

Table 9.17: Correction factors when $|p - \pi_0| > \frac{1}{2n}$

Condition	Correction Factor	Standardized Test Statistic
$p - \pi_0 > 0$	$-\frac{1}{2n}$	$z_{\text{obs}} = \frac{p - \pi_0 - \frac{1}{2n}}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}}$
$p - \pi_0 < 0$	$+\frac{1}{2n}$	$z_{\text{obs}} = \frac{p - \pi_0 + \frac{1}{2n}}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}}$

Example 9.17 ▷ *Normal Approximation: Graduates' Jobs* ◁ A recent report claimed that 20% of all college graduates find a job in their chosen field of study. A random sample of 500 graduates found that 90 obtained work in their field. Using a normal approximation to the distribution of P ,

- Is there statistical evidence to refute the claim at the $\alpha = 0.05$ level?
- Compute a 95% confidence interval for the true proportion of college graduates that find work in their chosen field of study using (8.44) on page 323.

Solution: The answers are as follows:

- Use the five-step procedure.

Step 1: **Hypotheses** — The null and alternative hypotheses to test whether or not 20% of college graduates find work in their chosen field are

$$H_0 : \pi = 0.20 \text{ versus } H_1 : \pi \neq 0.20.$$

Step 2: **Test Statistic** — The test statistic chosen is P , where P is the proportion of college graduates finding work in their chosen field. Provided H_0 is true,

$$P \sim N\left(\pi_0, \sqrt{\frac{\pi_0(1-\pi_0)}{n}}\right)$$

and the standardized test statistic is

$$Z = \frac{P - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \sim N(0, 1).$$

Step 3: **Rejection Region Calculations** — Because the standardized test statistic has an approximate $N(0, 1)$ distribution, and H_1 is a two-sided hypothesis, the rejection region is $|z_{\text{obs}}| > z_{0.975} = 1.96$. The value of the standardized test statistic is

Without Continuity Correction

$$\begin{aligned} z_{\text{obs}} &= \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \\ &= \frac{\frac{90}{500} - 0.2}{\sqrt{\frac{(0.2)(1-0.2)}{500}}} \\ &= -1.1180 \end{aligned}$$

With Continuity Correction

$$\begin{aligned} z_{\text{obs}} &= \frac{p - \pi_0 + \frac{1}{2n}}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \\ &= \frac{\frac{90}{500} - 0.2 + \frac{1}{1000}}{\sqrt{\frac{(0.2)(1-0.2)}{500}}} \\ &= -1.0621 \end{aligned}$$

OR

Step 4: **Statistical Conclusion** — The φ -value is $2 \cdot \mathbb{P}(Z \leq -1.118) = 0.2636$ or $2 \cdot \mathbb{P}(Z \leq -1.0621) = 0.2882$ for continuity corrections not used and used, respectively.

- I. From the rejection region, do not reject H_0 because $|z_{\text{obs}}| = |-1.1180|$ (no continuity correction) is not greater than 1.96, nor is $|z_{\text{obs}}| = |-1.0621|$ (continuity correction) greater than 1.96.
- II. From the φ -value, do not reject H_0 because the φ -value = 0.2636 (without continuity correction) or = 0.2882 (with continuity correction) is greater than 0.05.

Fail to reject H_0 .

The S code to calculate standardized test statistics and φ values is

```
> Y <- 90
> n <- 500
> p <- Y/n
> PI <- 0.2
> zobs <- (p - PI)/sqrt((PI * (1 - PI))/n)
> pval <- 2 * pnorm(zobs)
> zobsC <- (p - PI + 1/(2 * n))/sqrt((PI * (1 - PI))/n)
> pvalC <- 2 * pnorm(zobsC)
> round(c(zobs, pval, zobsC, pvalC), 4)
[1] -1.1180 0.2636 -1.0621 0.2882
```

Step 5: **English Conclusion** — There is not sufficient evidence to suggest the proportion of college graduates finding work in their chosen fields of study is something other than 20%.

(b) An approximate 95% confidence interval is [0.1488049, 0.2160747] without a continuity correction and [0.1478847, 0.2171388] with a continuity correction. One is 95% confident that the true proportion of college graduates finding work in their chosen fields of study lies in [0.1488049, 0.2160747]. Note that this confidence interval contains the hypothesized value of 0.20, corroborating the decision to fail to reject the null hypothesis.

The calculation of 95% confidence intervals as well as verifications of the calculated ϕ -values from step 4 are computed with `prop.test()` both without and with continuity corrections being used:

```
> prop.test(x=90, n=500, p=.2, correct=FALSE)

      1-sample proportions test without continuity correction

data:  90 out of 500, null probability 0.2
X-squared = 1.25, df = 1, p-value = 0.2636
alternative hypothesis: true p is not equal to 0.2
95 percent confidence interval:
 0.1488049 0.2160747
sample estimates:
      p
0.18

> prop.test(x=90, n=500, p=.2, correct=TRUE)

      1-sample proportions test with continuity correction

data:  90 out of 500, null probability 0.2
X-squared = 1.1281, df = 1, p-value = 0.2882
alternative hypothesis: true p is not equal to 0.2
95 percent confidence interval:
 0.1478847 0.2171388
sample estimates:
      p
0.18
```

Note that the output for `prop.test()` does not give a z_{obs} -value; rather it reports a χ_{obs}^2 -value, denoted **X-squared** in the **S** output, with one degree of freedom. Provided one uses the relationship $Z^2 = \chi_1^2$, it is possible to see that the z_{obs} -values reported in step 3 correspond to the **X-squared** values given in the **S** output from using `prop.test()` without and with continuity correction. That is, $-1.118034^2 = 1.25$ and $-1.062132^2 = 1.1281$.

Although the approximation procedures presented in this section lead to the same conclusion as the exact test in the previous section when applied to Example 9.16 on page 380, the approximation procedures of this section are only valid when applied to large samples. In contrast, the exact test presented in the last section will work for both large and small samples and is generally preferred over large sample approximation procedures when the user has access to a computer.

9.9.3 Testing Equality of Proportions with Fisher's Exact Test

One of the most common ways to present numerical data is in a table. When presented with a 2×2 table, where 2×2 refers to the dimensions of the number of internal cells, if the sample size is small, equality of proportions should be tested with **Fisher's exact test**. That is, $H_0 : \pi_X = \pi_Y$, where $X \sim \text{Bin}(m, \pi_X)$ and $Y \sim \text{Bin}(n, \pi_Y)$ are the numbers of successes observed from two independent binomial random variables. To compute Fisher's exact test, let $N = m + n$ be the total sample size and $k = x + y$ be the total number of observed successes. Table 9.18 shows the general form of such a table.

Table 9.18: General form of a 2×2 table

	Success	Failure	Total
X Sample	x	$m - x$	m
Y Sample	y	$n - y$	n
	k	$N - k$	N

Fisher's exact test uses the number of successes from the X sample as its test statistic, namely X . The observed value of X is denoted x . In performing the exact test, the total number of successes is considered fixed. That is, $x + y = k$ is a fixed quantity in the derivation of the test. Specifically,

$$\mathbb{P}(X = i | X + Y = k) = \frac{\binom{m}{i} \binom{n}{k-i}}{\binom{N}{k}}, \text{ where } i = \max\{0, k - n\}, \dots, \min\{m, k\}. \quad (9.10)$$

Note that (9.10) is a hypergeometric distribution, $\text{Hyper}(m, n, k)$, where the parameters are m , n , and k . As in Section 9.9.1, the three possible alternative hypotheses and the respective p -value calculation formulas are presented in Table 9.19 on the following page. Since the distribution of the statistic is obtained by constructing all possible 2×2 tables, the test has historically been used with small samples. With the advent of inexpensive computing power, it is now feasible to use Fisher's exact test on relatively large samples with fixed marginals.

A statistic that measures how associated X and Y are is the **odds ratio**. It is frequently used in biomedical and sociological studies to measure the association between two variables. The odds ratio is defined as

$$\theta = \frac{\pi_X / (1 - \pi_X)}{\pi_Y / (1 - \pi_Y)}. \quad (9.11)$$

An odds ratio other than 1 indicates there is a relationship between X and Y , while an odds ratio of exactly 1 indicates that X and Y are independent. If the odds ratio is larger than 1, π_X is greater than π_Y ; and if smaller, π_X is less than π_Y .

Only R computes a $(1 - \alpha) \cdot 100\%$ confidence interval for the odds ratio. R's procedure uses maximum likelihood techniques with the non-central hypergeometric distribution to compute the confidence interval. The procedure is beyond the scope of this text. When R's $(1 - \alpha) \cdot 100\%$ confidence interval for the odds ratio upper bound is less than 1, one can be $(1 - \alpha) \cdot 100\%$ confident that π_X is less than π_Y . Likewise, when the lower bound of a $(1 - \alpha) \cdot 100\%$ confidence interval for the odds ratio is greater than 1, one can be $(1 - \alpha) \cdot 100\%$ confident that π_X is greater than π_Y . R is capable of performing this test for one-sided alternative hypotheses; however, at the time of writing, S-PLUS was

Table 9.19: Summary for testing the proportion of successes with Fisher's exact test

Null Hypothesis — $H_0 : \pi_X = \pi_Y$ Test Statistic's Value — $x =$ number of observed successes from X sample

Alternative Hypothesis	\wp -Value Formula
$H_1 : \pi_X < \pi_Y$	$\mathbb{P}(X \leq x H_0) = \sum_{i=\max\{0, k-n\}}^x \frac{\binom{m}{i} \binom{n}{k-i}}{\binom{N}{k}}$
$H_1 : \pi_X > \pi_Y$	$\mathbb{P}(X \geq x H_0) = \sum_{i=x}^{\min\{m, k\}} \frac{\binom{m}{i} \binom{n}{k-i}}{\binom{N}{k}}$
$H_1 : \pi_X \neq \pi_Y$	$\sum_{i=\max\{0, k-n\}}^{\min\{m, k\}} I(\mathbb{P}(X = i) \leq \mathbb{P}(X = x)) \frac{\binom{m}{i} \binom{n}{k-i}}{\binom{N}{k}}$

not. Additionally, the current version of S-PLUS does not compute a confidence interval associated with `fisher.test()`.

Example 9.18 ▷ *Fisher's Exact Test: Delinquents in Glasses* ◁ A researcher wants to discover if the proportion of non-delinquent juveniles who wear glasses is different from that of juvenile delinquents. He collects the information found in Table 9.20 from juveniles who failed a vision test. Test whether the proportion of non-delinquents who wear glasses is different from the proportion of juvenile delinquents who wear glasses at an α level of 0.05.

Table 9.20: Juveniles who failed a vision test classified by delinquency and glasses wearing (Weindling et al., 1986)

	Wear Glasses	Do Not Wear Glasses	Totals
Juvenile Delinquents	1	8	9
Non-delinquents	5	2	7
Totals	6	10	16

Solution: To solve this problem, use Fisher's exact test and the five-step procedure.

Step 1: **Hypotheses** — The null and alternative hypotheses to test whether the proportion of non-delinquents who wear glasses is different from the proportion of juvenile delinquents who wear glasses are

$$H_0 : \pi_X = \pi_Y \text{ versus } H_1 : \pi_X \neq \pi_Y.$$

In this case, the random variable X will represent the number of juvenile delinquents who wear glasses, and the random variable Y will represent the number of non-delinquents who wear glasses.

Step 2: **Test Statistic** — The test statistic chosen is X , where X is the number of juvenile delinquents who wear glasses. The observed value of the test statistic is $x = 1$. Provided H_0 is true, and conditioning on the fact that $X + Y = k$, $X \sim \text{Hyper}(m, n, k)$.

Step 3: **Rejection Region Calculations** — Rejection is based on the φ -value, so none are required.

Step 4: **Statistical Conclusion** — To compute the φ -value, compute

$$\sum_{i=\max\{0, k-n\}}^{\min\{m, k\}} I(\mathbb{P}(X = i) \leq \mathbb{P}(X = x)) \frac{\binom{m}{i} \binom{n}{k-i}}{\binom{N}{k}} = \sum_{i=\max\{0, 6-7\}}^{\min\{9, 6\}} I(\mathbb{P}(X = i) \leq \mathbb{P}(X = 1)) \frac{\binom{9}{i} \binom{7}{6-i}}{\binom{16}{6}} = 0.035$$

For such a small sample, the seven possible 2×2 tables that can be constructed where $k = 6$, $m = 9$, and $n = 7$ and their respective φ -values are shown in Table 9.21. Since the φ -value is 0.035, one rejects H_0 because 0.035 is less than 0.05.

Reject H_0 .

Table 9.21: Seven possible 2×2 tables that can be constructed where $k = 6$, $m = 9$, and $n = 7$, with their associated probabilities

Table	Probability	Table	Probability	Table	Probability	Table	Probability
0 9 6 1	0.00087	1 8 5 2	0.0236	2 7 4 3	0.15734	3 6 3 4	0.36713
4 5 2 5	0.33042	5 4 1 6	0.11014	6 3 0 7	0.0104		

Step 5: **English Conclusion** — There is sufficient evidence to suggest the proportion of non-delinquents who wear glasses is different from the proportion of juvenile delinquents who wear glasses.

The S code to enter the information from Table 9.20 on the preceding page and to perform Fisher's exact test is

```
> JD <- matrix(c(1,5,8,2), nrow=2,
+             dimnames=list>Youth=c("Delinquent", "Non-delinquent"),
+             Glasses=c("Yes", "No")))
> JD # Output is for R, S-PLUS is slightly different.
```

Youth	Glasses	
	Yes	No
Delinquent	1	8
Non-delinquent	5	2


```

> p <- dhyper(0:6,9,7,6) # Probabilities for the 7 possible 2X2 tables

> round(p,5)
[1] 0.00087 0.02360 0.15734 0.36713 0.33042 0.11014 0.01049
> pobs <- dhyper(1,9,7,6)
> pval <- sum(p[p<=pobs])

> pval
[1] 0.03496503
> fisher.test(JD)          # Output is for R, S-PLUS is slightly different.

```

Fisher's Exact Test for Count Data

```

data: JD
p-value = 0.03497
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
0.0009525702 0.9912282442
sample estimates:
odds ratio
0.06464255

```

Note that values of θ farther from 1.0 in a given direction represent stronger levels of association ($0 < \theta < \infty$). In this case, $\theta = 0.06$ means that the odds ratio for non-delinquents wearing glasses is $1/0.06 = 15.5$ times the odds ratio for delinquents wearing glasses. This is a very strong association. ■

Example 9.19 ▷ *Fisher's exact test of $\pi_X = \pi_Y$: Heart Attacks* ◁ Physicians want to know if taking aspirin will help them avoid heart attacks. A group collects the information found in Table 9.22. Help them test to see if taking aspirin is beneficial in the prevention of heart attacks at an α level of 0.05.

Table 9.22: Observed heart attacks for those physicians taking aspirin and a placebo (Hennekens, 1988)

	Heart Attack	No Heart Attack	Totals	
Aspirin	104	10,933	11,037	= m
Placebo	189	10,845	11,034	= n
Totals	293 = k	21,778	22,071	= N

Solution: To solve this problem, use Fisher's exact test and the five-step procedure.

Step 1: **Hypotheses** — The null and alternative hypotheses to test whether the proportion of physicians who suffer heart attacks while taking aspirin is less than the proportion of physicians who suffer heart attacks while taking a placebo are

$$H_0 : \pi_X = \pi_Y \text{ versus } H_1 : \pi_X < \pi_Y.$$

In this case, let the random variable X represent the number of physicians who had a heart attack while taking aspirin, and let the random variable Y represent the number of physicians who had a heart attack while taking a placebo.

Step 2: **Test Statistic** — The test statistic chosen is X , where X is the number of physicians who had a heart attack while taking aspirin. Provided H_0 is true, and conditioning on the fact that $X + Y = k$, $X \sim \text{Hyper}(m, n, k)$. The observed value of the test statistic is $x = 104$. To enter the data from Table 9.22 on the preceding page into S, type the following code:

```
> HA <- matrix(c(104,189,10933,10845), nrow=2,
+             dimnames=list(Treatment=c("Aspirin","Placebo"),
+             Outcome=c("Heart attack","No heart attack")))
> HA # Output is for R, S-PLUS is slightly different
      Outcome
Treatment Heart attack No heart attack
Aspirin      104      10933
Placebo      189      10845
```

Step 3: **Rejection Region Calculations** — Rejection is based on the ϕ -value, so none are required.

Step 4: **Statistical Conclusion** — To calculate the ϕ -value, compute

$$\mathbb{P}(X \leq x_{\text{obs}} = 104) = \sum_{i=0}^{104} \frac{\binom{104+10933}{i} \binom{189+10845}{293-i}}{\binom{22071}{293}}$$

Note that the limits on the sum are typically the $\max\{0, x - n\}$ and x . In this case $x - n = 104 - 11,034 = -10,930$, so the lower limit of the sum will be zero. This calculation should be done with a computer, so the S code to do so follows. Note that the data from the table must have been entered as shown in step 2.

```
> pval <- phyper(104,104+10933,189+10845,104+189) # x, m, n, k
> pval
[1] 3.252711e-07
> fisher.test(HA, alternative="less") # Only in R
```

Fisher's Exact Test for Count Data

```
data: HA
p-value = 3.253e-07
alternative hypothesis: true odds ratio is less than 1
95 percent confidence interval:
 0.0000000 0.6721508
sample estimates:
odds ratio
 0.5458537
```

Because the ϕ -value is approximately 0, which is less than 0.05, reject H_0 .

Reject H_0 .

Step 5: **English Conclusion** — There is sufficient evidence to suggest taking aspirin is beneficial in the prevention of heart attacks for physicians at an α level of 0.05.

Note that the odds ratio for physicians having a heart attack taking a placebo is $1/0.546 = 1.83$ times the odds ratio for physicians who take aspirin. ■

9.9.4 Large Sample Approximation for Testing the Difference of Two Proportions

In Section 9.9.3, Fisher's exact test was presented for testing the equality of proportions for two independent random samples taken from Bernoulli populations of sizes m and n , respectively. Once the sample sizes become large for Fisher's exact test, even computers begin to have difficulties. Thus, there exists a procedure for approximating the distribution of $P_X - P_Y$ that will lead to a test that does not have nearly the computational requirements of Fisher's exact test. In Section 8.4.2, it was argued that

$$P_X - P_Y \sim N\left(\pi_X - \pi_Y, \sqrt{\frac{\pi_X(1-\pi_X)}{m} + \frac{\pi_Y(1-\pi_Y)}{n}}\right) \quad (9.12)$$

when taking independent random samples of sizes m and n from $Bernoulli(\pi_X)$ and $Bernoulli(\pi_Y)$ populations, respectively. Using (9.12),

$$Z = \frac{(P_X - P_Y) - (\pi_X - \pi_Y)}{\sqrt{\frac{\pi_X(1-\pi_X)}{m} + \frac{\pi_Y(1-\pi_Y)}{n}}} \sim N(0, 1). \quad (9.13)$$

Unfortunately, the values of π_X and π_Y are unknown. In Section 8.4.2, π_X and π_Y were replaced with their maximum likelihood estimators, $\hat{\pi}_X = P_X$ and $\hat{\pi}_Y = P_Y$, respectively, to create the asymptotic confidence interval in (8.48). To create a standardized test statistic with an approximate normal distribution, the same approach will be used. That is,

$$Z = \frac{(P_X - P_Y) - \delta_0}{\sqrt{\frac{P_X(1-P_X)}{m} + \frac{P_Y(1-P_Y)}{n}}} \sim N(0, 1) \quad (9.14)$$

can be used to test the null hypothesis $H_0 : \pi_X - \pi_Y = \delta_0$. It is often the case that δ_0 is zero. In this case, it is standard practice to create a pooled estimate of the population proportions such that $\pi_X = \pi_Y = \pi$. The pooled estimate of π , denoted P , is

$$P = \frac{X + Y}{m + n} \quad (9.15)$$

which is simply an estimate of the total proportion of successes. When this estimate is used, the standardized test statistic becomes

$$Z = \frac{(P_X - P_Y)}{\sqrt{P(1-P)\left(\frac{1}{m} + \frac{1}{n}\right)}} \sim N(0, 1). \quad (9.16)$$

There are advantages and disadvantages to both (9.14) and (9.16) as test statistics. The S function `prop.test()` bases its confidence interval construction on (9.14) and uses (9.16) for testing hypotheses. Table 9.23 on the facing page uses the standardized test statistic in (9.16) and provides the rejection regions for the three possible alternative hypotheses.

Table 9.23: Summary for testing the differences of the proportions of successes in two binomial experiments (large sample approximation)

Null Hypothesis — $H_0 : \pi_X = \pi_Y$

Standardized Test
Statistic's Value — $z_{\text{obs}} = \frac{p_X - p_Y}{\sqrt{p(1-p)\left(\frac{1}{m} + \frac{1}{n}\right)}}$

Alternative Hypothesis	Rejection Region
$H_1 : \pi_X < \pi_Y$	$z_{\text{obs}} < z_\alpha$
$H_1 : \pi_X > \pi_Y$	$z_{\text{obs}} > z_{1-\alpha}$
$H_1 : \pi_X \neq \pi_Y$	$ z_{\text{obs}} > z_{1-\alpha/2}$
When $ p_X - p_Y > \frac{1}{2} \left(\frac{1}{m} + \frac{1}{n}\right)$, use a correction factor as in Table 9.24.	

When $|p_X - p_Y| > \frac{1}{2} \left(\frac{1}{m} + \frac{1}{n}\right)$, some statisticians advocate using a continuity correction when calculating confidence intervals and standardized test statistics' values. A continuity correction of $\pm \frac{1}{2} \left(\frac{1}{m} + \frac{1}{n}\right)$ is automatically applied when using the `S` function `prop.test()` on two samples. The continuity corrections that are applied, as well as the standardized test statistic calculations, can be found in Table 9.24. When applying the continuity correction to (8.48) on page 328, recall that the continuity correction is subtracted and added to the lower and upper confidence limits, respectively.

Table 9.24: Correction factors when $|p_X - p_Y| > \frac{1}{2} \left(\frac{1}{m} + \frac{1}{n}\right)$

Condition	Correction Factor	Standardized Test Statistic
$p_X - p_Y > 0$	$-\frac{1}{2} \left(\frac{1}{m} + \frac{1}{n}\right)$	$z_{\text{obs}} = \frac{p_X - p_Y - \frac{1}{2} \left(\frac{1}{m} + \frac{1}{n}\right)}{\sqrt{p(1-p) \left(\frac{1}{m} + \frac{1}{n}\right)}}$
$p_X - p_Y < 0$	$+\frac{1}{2} \left(\frac{1}{m} + \frac{1}{n}\right)$	$z_{\text{obs}} = \frac{p_X - p_Y + \frac{1}{2} \left(\frac{1}{m} + \frac{1}{n}\right)}{\sqrt{p(1-p) \left(\frac{1}{m} + \frac{1}{n}\right)}}$

Example 9.20 ▷ *Large Sample Test of $\pi_X = \pi_Y$: Heart Attacks* ◁ Use the data from Table 9.22 on page 390 to test whether physicians who take aspirin are less likely to suffer heart attacks than those who take a placebo at an α level of 0.05. Base the test on the large sample approximation procedures found in Table 9.23.

Solution: To solve this problem, use the five-step procedure.

Step 1: **Hypotheses** — The null and alternative hypotheses to test whether the proportion of physicians who suffer heart attacks while taking aspirin is less than the proportion

of physicians who suffer heart attacks while taking a placebo are

$$H_0 : \pi_X = \pi_Y \text{ versus } H_1 : \pi_X < \pi_Y.$$

In this case, let the random variable X represent the number of physicians who had a heart attack while taking aspirin, and let the random variable Y represent the number of physicians who had a heart attack while taking a placebo.

Step 2: **Test Statistic** — The test statistic chosen is $P_X - P_Y$ since $E[P_X - P_Y] = \pi_X - \pi_Y$. The standardized test statistic under the assumption that H_0 is true is

$$Z = \frac{P_X - P_Y}{\sqrt{P(1-P)\left(\frac{1}{m} + \frac{1}{n}\right)}}$$

Step 3: **Rejection Region Calculations** — Because the standardized test statistic has an approximate $N(0, 1)$ distribution and H_1 is a lower one-sided hypothesis, the rejection region is $z_{\text{obs}} < z_{0.05} = -1.645$. The pooled estimate of π is $p = \frac{x+y}{m+n} = \frac{293}{22071}$. The value of the standardized test statistic is

Without Continuity Correction

$$\begin{aligned} z_{\text{obs}} &= \frac{p_X - p_Y}{\sqrt{p(1-p)\left(\frac{1}{m} + \frac{1}{n}\right)}} \\ &= \frac{\frac{104}{11037} - \frac{189}{11034}}{\sqrt{\frac{293}{22071}\left(1 - \frac{293}{22071}\right)\left(\frac{1}{11037} + \frac{1}{11034}\right)}} \\ &= -5.01139 \end{aligned}$$

OR

With Continuity Correction

$$\begin{aligned} z_{\text{obs}} &= \frac{p_X - p_Y + \frac{1}{2}\left(\frac{1}{m} + \frac{1}{n}\right)}{\sqrt{p(1-p)\left(\frac{1}{m} + \frac{1}{n}\right)}} \\ &= \frac{\frac{104}{11037} - \frac{189}{11034} + \frac{1}{2}\left(\frac{1}{11037} + \frac{1}{11034}\right)}{\sqrt{\frac{293}{22071}\left(1 - \frac{293}{22071}\right)\left(\frac{1}{11037} + \frac{1}{11034}\right)}} \\ &= -4.94258 \end{aligned}$$

Step 4: **Statistical Conclusion** — The φ -value is $\mathbb{P}(Z \leq z_{\text{obs}})$ and is approximately 0 for both cases. This is less than 0.05, so reject H_0 .

Reject H_0 .

Step 5: **English Conclusion** — There is sufficient evidence to suggest taking aspirin is beneficial in the prevention of heart attacks for physicians at an α level of 0.05.

To perform this test with S, enter

```
> prop.test(c(104,189), c(11037,11034), correct=FALSE)
```

```
2-sample test for equality of proportions without continuity
correction
```

```
data: c(104, 189) out of c(11037, 11034)
X-squared = 25.0139, df = 1, p-value = 5.692e-07
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.010724297 -0.004687751
sample estimates:
  prop 1    prop 2 
0.00942285 0.01712887
```

```
> prop.test(c(104,189), c(11037,11034), correct=TRUE)
```

```
2-sample test for equality of proportions with continuity
correction
```

```
data: c(104, 189) out of c(11037, 11034)
X-squared = 24.4291, df = 1, p-value = 7.71e-07
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.010814914 -0.004597134
sample estimates:
  prop 1    prop 2 
0.00942285 0.01712887
```

Notice that if z_{obs} is squared, it will be equal to the values of X-squared in the S output. ■

9.10 Problems

1. Define α and β for a test of hypothesis. What is the quantity $1 - \beta$ called?
2. How can β be made small in a given hypothesis test with fixed α ?
3. Using a 5% significance level, what is the power of the test $H_0 : \mu = 100$ versus $H_1 : \mu \neq 100$ if the true standard deviation is $\sigma = 50$ grams?
4. An experiment was conducted to investigate how the resistance of rubber to abrasion is affected by the hardness of the rubber and its tensile strength. The data come from Hand et al. (1994, Data Set #6 Abrasion Loss) and are stored in the data frame **Rubber** of the **MASS** package. The abrasion loss is measured in grams/hour; the hardness, in degrees shore; and the tensile strength, in kg/cm². Use the five-step procedure to test whether $H_0 : \mu = 170$ versus $H_1 : \mu < 170$ for abrasion loss (`loss`).
5. An apartment appraiser in Vitoria, Spain, feels confident in his appraisals of 90m² or larger pisos (apartments) provided his variability is less than 60,000€². Due to constant movement in the housing market, the regional housing authority suspects the appraiser's variability may be greater than 60,000€². Is there evidence to support the suspicions of the regional housing authority? Test the appropriate hypothesis at the 5% significance level using the five-step procedure. The appraised values of apartments in Vitoria are stored in the variable `totalprice` of the `vit2005` data frame.
6. The Hubble Space Telescope was put into orbit on April 25, 1990. Unfortunately, on June 25, 1990, a spherical aberration was discovered in Hubble's primary mirror. To correct this, astronauts had to work in space. To prepare for the mission, two teams of astronauts practiced making repairs under simulated space conditions. Each team of astronauts went through 15 identical scenarios. The times to complete each scenario were recorded in days. Is one team better than the other? If not, can both teams complete the mission in less than 3 days? Use a 5% significance level for all tests. The data are stored in the data frame **Hubble**.
7. The research and development department of an appliance company suspects the energy consumption required of their 18 cubic foot refrigerator can be reduced by a slight modification to the current motor. Thirty 18 cubic foot refrigerators were randomly selected from the company's warehouse. The first 15 had their motors modified while the last 15 were left intact. The energy consumption (kilowatts) for a 24 hour period for each refrigerator was recorded and stored in the data frame **Refrigerator**. The refrigerators with the design modification are stored in the variable `modelA` and those without the design modification are stored in the variable `modelB`. Is there evidence that the design modification reduces the refrigerators' average energy consumption?
8. The Yonalasee tennis club has two systems to measure the speed of a tennis ball. The local tennis pros suspects one system (`Speed1`) consistently records faster speeds. To test her suspicions, she sets up both systems and records the speeds of 12 serves (three serves from each side of the court). The values are stored in the data frame **Tennis** in the variables `Speed1` and `Speed2`. The recorded speeds are in kilometers per hour. Does the evidence support the tennis pro's suspicion? Use $\alpha = 0.10$.
9. An advertising agency is interested in targeting the appropriate gender for a new "low-fat" yogurt. In a national survey of 1200 women, 825 picked the "low-fat" yogurt over

a regular yogurt. Meanwhile, 525 out of 1150 men picked the “low-fat” yogurt over the regular yogurt. Given these results, should the advertisements be targeted at a specific gender? Test the appropriate hypothesis at the $\alpha = 0.01$ level.

10. A plastics manufacturer makes two sizes of milk containers: half-gallon and gallon sizes. The time required for each size to dry is recorded in seconds in the data frame **MilkCarton**. Test to see if there are differences in average drying times between the container sizes.
11. A multinational conglomerate has two textile centers in two different cities. In order to make a profit, each location must produce more than 1000 kilograms of refined wool per day. A random sample of the wool production in kilograms on five different days over the last year for the two locations was taken. The results are stored in the data frame **Wool**. Based on the collected data, does the evidence suggest the locations are profitable? Is one location superior to the other?
12. Use the data frame **Fertilize**, which contains the height in inches for self-fertilized plants in the variable **self** to
 - (a) Test if the data suggest that the average height of self-fertilized plants is more than 17 inches. (Use $\alpha = 0.05$.)
 - (b) Compute a one-sided 95% confidence interval for the average height of self-fertilized plants ($H_1 : \mu > 17$).
 - (c) Compute the required sample size to obtain a power of 0.90 if $\mu_1 = 18$ inches assuming that $\sigma = s$.
 - (d) What is the power of the test in part (a) if $\sigma = s$ and $\mu_1 = 18$.
13. A manufacturer of lithium batteries has two production facilities. One facility manufactures a battery with an advertised life of 180 hours (**facilityA**), while the second facility manufactures a battery with an advertised life of 200 hours (**facilityB**). Both facilities are trying to reduce the variance in their products’ lifetimes. Is the variability in battery life equivalent, or does the evidence suggest the facility producing 180 hour batteries has smaller variability than the facility producing 200 hour batteries? Use the data frame **Battery** with $\alpha = 0.05$ to test the appropriate hypothesis.
14. In the construction of a safety strobe, a particular manufacturer can purchase LED diodes from one of two suppliers. It is critical that the purchased diodes conform to their stated specifications with respect to diameter since they must be mated with a fixed width cable. The diameter in millimeters for a random sample of 15 diodes from each of the two suppliers is stored in the data frame **Leddiode**. Based on the data, is there evidence to suggest a difference in variabilities between the two suppliers? Use an α level of 0.01.
15. The technology at a certain computer manufacturing plant allows silicon sheets to be split into chips using two different techniques. In an effort to decide which technique is superior, 28 silicon sheets are randomly selected from the warehouse. The two techniques of splitting the chips are randomly assigned to the 28 sheets so that each technique is applied to 14 sheets. The results from the experiment are stored in the data frame **Chips**. Use $\alpha = 0.05$, and test the appropriate hypothesis to see if there are differences between the two techniques. The values recorded in **Chips** are the number of usable chips from each silicon sheet.

16. Phenylketonuria (PKU) is a genetic disorder that is characterized by an inability of the body to utilize the essential amino acid, phenylalanine. Research suggests patients with phenylketonuria have deficiencies in coenzyme Q10. The data frame **Pheny1** records the level of Q10 at four different times for 46 patients diagnosed with PKU. The variable **Q10.1** contains the level of Q10 measured in μM for the 46 patients. **Q10.2**, **Q10.3**, and **Q10.4** record the values recorded at later times, respectively, for the 46 patients (Artuch et al., 2004).
- Normal patients have a Q10 reading of $0.69 \mu\text{M}$. Using the variable **Q10.2**, is there evidence that the mean value of Q10 in patients diagnosed with PKU is less than $0.69 \mu\text{M}$? (Use $\alpha = 0.01$.)
 - Patients diagnosed with PKU are placed on strict vegetarian diets. Some have speculated that patients diagnosed with PKU have low Q10 readings because meats are rich in Q10. Is there evidence that the patients' Q10 level decreases over time? Construct a 99% confidence interval for the means of the Q10 levels using **Q10.1** and **Q10.4**
17. According to the Pamplona, Spain, registration, 0.4% of immigrants in 2002 were from Bolivia. In June of 2005, a sample of 3740 registered foreigners was randomly selected. Of these, 87 were Bolivians. Is there evidence to suggest immigration from Bolivia has increased? (Use $\alpha = 0.05$.)
18. Find the power for the hypothesis $H_0 : \mu = 65$ versus $H_1 : \mu > 65$ if $\mu_1 = 70$ at the $\alpha = 0.01$ level assuming $\sigma = s$ for the variable **hard** in the data frame **Rubber** of the **MASS** package.
19. The director of urban housing in Vitoria, Spain, claims that at least 50% of all apartments have more than one bathroom and that at least 75% of all apartments have an elevator.
- Can the director's claim about bathrooms be contradicted? Test the appropriate hypothesis using $\alpha = 0.10$. Note that the number of bathrooms is stored in the variable **toilets** in the data frame **vit2005**.
 - Can the director's claim about elevators be substantiated using an α level of 0.10? Use both an approximate method as well as an exact method to reach a conclusion. Are the methods in agreement?
 - Test whether the proportion of apartments built prior to 1980 without garages have a higher proportion with elevators than without elevators.
20. A rule of thumb used by realtors in Vitoria, Spain, is that each square meter will cost roughly €3000. However, there is some suspicion that this figure is high for apartments in the 55 to 66 m^2 range. Use a 5 m^2 bracket, that is, [55, 60) and [60, 65), to see if evidence exists that the average difference between the larger and smaller apartment sizes is less than €15,000.
- Use the data frame **vit2005** and the variables **totalprice** and **area** to test the appropriate hypothesis at a 5% significance level.
 - Are the assumptions for using a t -test satisfied? Explain.
 - Does the answer for (b) differ if the variances are assumed to be equal? Can the hypothesis of equal variances be rejected?

21. A survey was administered during the first trimester of 2005 in the Spanish province of Navarra. The numbers of unemployed people according to urban and rural areas and gender follow.

Unemployment in Navarra, Spain, in 2005

	Male	Female	Totals
Urban	4734	6161	10895
Rural	3259	4033	7292
Totals	4933	10194	18127

- (a) Test to see if there is evidence to suggest that $\pi_{\text{urban.male}} < \pi_{\text{urban.female}}$ at $\alpha = 0.05$.
- (b) Use an exact test to see if the evidence suggests $\pi_{\text{urban.female}} > 0.55$.
- (c) Is there evidence to suggest the unemployment rate for rural females is greater than 50%? Use $\alpha = 0.05$ with an exact test to reach a conclusion.
- (d) Does evidence suggest that $\pi_{\text{urban.female}} > \pi_{\text{rural.female}}$?
22. The owner of a transportation fleet is evaluating insurance policies for transporting hazardous waste. The owner has narrowed his possibility of insurers to two companies (A and B). Insurance company A claims to have the least expensive policies on the market while insurer B disputes the claim. To evaluate company A's claim, the owner requests the last 100 policies issued by each insurer. The means and standard deviations are €325 and €85 for company A and €340 and €75 for company B. Based on these summary statistics, the owner was not convinced that company A actually had less expensive rates. Consequently, a representative from company A was sent to speak to the owner. The representative from company A convinced the owner to take another look at the numbers. This time, insurance quotes were sought from both insurers for the next 15 transportation of hazardous waste jobs. Results are given in the data frame **InsurQuotes**. Analyze these data. How is it possible the owner changed his mind with a sample of size 15 versus the results based on a sample of size 100?
23. The data frame **vit2005** contains housing information for the Spanish city Vitoria collected in 2005. Use the variables **age** and **garage** to see if the proportion of abodes with garages has increased since 1980. Use $\alpha = 0.05$ to reach a conclusion.
- (a) Use Fisher's exact test to test the appropriate hypothesis.
- (b) Use a normal approximation to test the same hypothesis as was tested in (a).
- (c) Compute the ϕ -value for (a) using the hypergeometric distribution.
- (d) Compute an exact and an approximate 95% confidence interval for the proportion of abodes with garages built after 1980.
24. Environmental monitoring is done in many fashions, including tracking levels of different chemicals in the air, underground water, soil, fish, milk, and so on. It is believed that milk cows eating in pastures where gamma radiation from iodine exceeds $0.3 \mu\text{Gy/h}$ in turn leads to milk with iodine concentrations in excess of 3.7 MBq/m^3 . Assuming the distribution of iodine in pastures follows a normal distribution with a standard deviation of $0.015 \mu\text{Gy/h}$, determine the required sample size to detect a 2% increase in baseline gamma radiation ($0.3\mu\text{Gy/h}$) using an $\alpha = 0.05$ significance level with probability 0.99 or more.

25. A local farmer packages and freezes his spinach. He claims that the packages weigh 340 grams and have a standard deviation of no more than $\sigma = 15$ grams. The manager of a local organic supermarket is somewhat skeptical of the farmer's claim and decides to test the claim using a random sample of 10 frozen spinach packages.
- Find the critical region of the test if $\alpha = 0.05$.
 - Find the power of the test if $\sigma = 10$.
26. A cell phone provider has estimated that it needs revenues of €2 million per day in order to make a profit and remain in the market. If revenues are less than €2 million per day, the company will go bankrupt. Likewise, revenues greater than €2 million per day cannot be handled without increasing staff. Assume that revenues follow a normal distribution with $\sigma = \text{€}0.5$ million and a mean of μ .
- Graphically depict the power function for testing $H_0 : \mu = 2$ versus $H_1 : \mu \neq 2$ if $n = 150$ and $\alpha = 0.05$ for values of μ ranging from 1.8 to 2.2.
 - Graphically depict the power for testing $H_0 : \mu = 2$ versus $H_1 : \mu \neq 2$ when $\mu_1 = 2.1$ and $n = 150$ for values of α ranging from 0.01 to 0.5.
 - Graphically depict the power for testing $H_0 : \mu = 2$ versus $H_1 : \mu \neq 2$ when $\mu_1 = 2.1$ and $\alpha = 0.05$ for values of n ranging from 1 to 365.
 - Generalize what is seen in the graphs for (a), (b), and (c).
27. Use simulation to compute the empirical significance level by generating 10,000 samples of size n from a $N(100, 28)$ population using $\alpha = 0.05$ to test the alternative hypothesis $H_1 : \mu \neq 100$. Use the command `set.seed(33)` so the answers can be reproduced.
- Use samples of size $n = 49$.
 - Use samples of size $n = 196$.
 - Use samples of size $n = 1936$.
 - Does increasing the sample size affect the significance level?
28. Use simulation to compute the empirical power for testing $H_0 : \mu = 100$ versus $H_1 : \mu > 100$ when $\mu = 108$ and sampling from a $N(100, 28)$ distribution. Use 10,000 samples with $n = 49$ in the simulation and `set.seed(14)` so that the results will be reproducible.
- Use a significance level of $\alpha = 0.05$.
 - Use a significance level of $\alpha = 0.20$.
 - Compute the theoretical power for scenarios (a) and (b). How do these values compare to those from the simulations.
 - What happens to the empirical power as α increases?
29. Test the null hypothesis that the mean life for a certain brand of 19 mm tubular tires is 1000 miles against the alternative hypothesis that it is less than 1000 miles. Assume that tubular tire life follows a normal distribution with $\sigma = 100$ miles.
- Find the probability of a type I error for $n = 16$ if the null hypothesis is rejected when the sample mean is less than or equal to 960 miles.
 - Plot the power function for $n = 16$ for values of μ between 900 and 1000 miles.
30. Given a normal population with unknown mean and a known variance of $\sigma^2 = 4$, test the hypothesis $H_0 : \mu = 10$ versus $H_1 : \mu < 10$ at the $\alpha = 0.05$ significance level.

- (a) Use the command `set.seed(28)` to generate 10,000 samples of size $n = 16$ from a $N(10, 2)$ population. Is the empirical significance level close to 5%?
- (b) Compute a 95% confidence interval for α when simulating 10,000 samples of size $n = 16$ from a $N(10, 2)$ population.
- (c) What is the theoretical power if $\mu_1 = 9.5$ for the given hypothesis test?
- (d) Graphically represent the power for testing $H_0 : \mu = 10$ versus $H_1 : \mu < 10$ for samples of size $n = 16$ from a $N(10, 2)$ population when $\alpha = 0.05$ for values of μ from 8 to 10.
- (e) Graphically represent the power for testing $H_0 : \mu = 10$ versus $H_1 : \mu < 10$ for samples of size $n = 16$ when $\mu_1 = 9.5$ for values of α ranging from 0.01 to 0.3.
31. Generate 10,000 samples of size $n_x = 20$ from $X \sim N(8, 2)$ and 10,000 samples of size $n_y = 20$ from $Y \sim N(6, 2)$. Use `set.seed(59)` so that the answers are reproducible. Assuming X and Y are independent and a 5% significance level,
- (a) What type of distribution does the statistic s_x^2/s_y^2 follow?
- (b) Create a density histogram of the 10,000 values of s_x^2/s_y^2 . Superimpose the theoretical sampling distribution of s_x^2/s_y^2 on the density histogram.
- (c) Compute the empirical significance level for testing $H_0 : \sigma_x^2/\sigma_y^2 = 1$ versus $H_1 : \sigma_x^2/\sigma_y^2 \neq 1$.
- (d) Plot the power function for testing $H_0 : \sigma_x^2/\sigma_y^2 = 1$ versus $H_1 : \sigma_x^2/\sigma_y^2 \neq 1$ for ratios of σ_x^2/σ_y^2 from 0.25 to 4.
- (e) Repeat (d) for $n_x = n_y = 100$.
- (f) Repeat (d) for $n_x = n_y = 200$.
- (g) Put the graphs from (d), (e), and (f) on the same graph.
- (h) Plot the power function for testing $H_0 : \sigma_x^2/\sigma_y^2 = 1$ versus $H_1 : \sigma_x^2/\sigma_y^2 \neq 1$ for α values between 0.01 and 0.5 if $\sigma_x^2/\sigma_y^2 = 2$ and
- (1) $n_x = n_y = 20$,
 - (2) $n_x = n_y = 100$, and
 - (3) $n_x = n_y = 200$.
- (i) Simulate the power for testing $H_0 : \sigma_x^2/\sigma_y^2 = 1$ versus $H_1 : \sigma_x^2/\sigma_y^2 \neq 1$ at the $\alpha = 0.5$ level when $\sigma_x^2/\sigma_y^2 = 2$ and
- (1) $n_x = n_y = 20$,
 - (2) $n_x = n_y = 100$, and
 - (3) $n_x = n_y = 200$.
 - (4) Compute the theoretical power for the three previous scenarios.

Chapter 10

Nonparametric Methods

10.1 Introduction

The statistical inference techniques presented in Chapter 8 and Chapter 9 are based on complete satisfaction of all of the assumptions made in the derivations of their sampling distributions. Indeed, many of the techniques in Chapters 8 and 9 are commonly referred to as parametric techniques since not only was the form of the underlying population (generally normal) stated, but so was one or more of the underlying distribution's parameters. This chapter introduces both distribution free tests as well as nonparametric tests. The collection of inferential techniques known as **distribution free** are based on functions of the sample observations whose corresponding random variable has a distribution that is independent of the specific distribution function from which the sample was drawn. Consequently, assumptions with respect to the underlying population are not required. **Nonparametric tests** involve tests of a hypothesis where there is no statement about the distribution's parameters. However, it is common practice to refer collectively to both distribution free tests and nonparametric tests simply as **nonparametric methods**.

When there are analogous parametric and nonparametric tests, comparisons between tests can be made based on power. The **power efficiency** of a test A relative to a test B is the ratio of n_b/n_a , where n_a is the number of observations required by test A for the power of test A to equal the power of test B when n_b observations are used. Since the power value is conditional on the type of alternative hypothesis and on the significance level, power efficiency can be difficult to interpret. One way to avoid this problem is to use the **asymptotic relative efficiency (ARE)** (a limiting power efficiency) of consistent tests. A test is consistent for a specified alternative if the power of the test when that alternative is true approaches 1 as the sample size approaches infinity (Gibbons, 1997). Provided that A and B are consistent tests of a null hypothesis H_0 and alternative hypothesis H_1 at significance level α , the asymptotic relative efficiency of test A to test B is the limiting ratio n_b/n_a , where n_a is the number of observations required by test A for the power of test A to equal the power of test B based on n_b observations while simultaneously $n_b \rightarrow \infty$ and $H_1 \rightarrow H_0$. Although the *ARE* considers infinite (hence not obtainable) sample sizes, the *ARE* provides a good approximation to the relative efficiency for many situations of practical interest. When a nonparametric technique has a parametric analog, the *ARE* will be used to compare the two techniques.

10.2 Sign Test

When the parent distribution is skewed or has long tails, the median is generally a better measure of the distribution's center than is the mean. In this section, a procedure

for testing hypotheses concerning the population median is introduced. This procedure is known as the **sign test**. A corresponding confidence interval formula for the median will also be derived.

To use the sign test, assume X_1, X_2, \dots, X_n is a random sample of n observations drawn from a continuous population with unknown median ψ . The sign test statistic, S , is defined as the number of positive differences among the $X_1 - \psi_0, X_2 - \psi_0, \dots, X_n - \psi_0$, where ψ_0 is the median from the null hypothesis $H_0 : \psi = \psi_0$. The distribution of S when H_0 is true is $S \sim \text{Bin}(n, \pi = 0.5)$.

The sign test may also be used for testing whether the median difference (ψ_D) between two dependent populations (X and Y) is equal to some value, $H_0 : \psi_D = \psi_0$. It is important to point out that ψ_D is not usually equal to $\psi_X - \psi_Y$. The only instance where ψ_D is equal to $\psi_X - \psi_Y$ is when X and Y are symmetric populations. For dependent samples, S is defined as the number of positive differences among the $X_1 - Y_1 - \psi_0, X_2 - Y_2 - \psi_0, \dots, X_n - Y_n - \psi_0$.

Since the assumption of a continuous population is a requirement for using the sign test, theoretically, there should not be any values that are exactly equal to the parameter being tested in the sample. However, due to rounding or crude measurements, it is not uncommon to observe sample values equal to ψ_0 , the value of ψ or ψ_D under the null hypothesis. There are several strategies one can pursue in dealing with values that are equal to the parameter being tested. Some of these include randomization, midranks, average statistic, average probability, least favorable statistic, range of probability, and omission of tied observations. The book by Pratt and Gibbons (1981) gives a more complete discussion of these various techniques. The final approach, **omission of tied observations**, consists of eliminating the value(s) in the sample that are equal to the parameter being tested. This is the approach that will be used in this text. This method leads to some loss of information; however, if the number of observations to be omitted is small compared to the sample size, this loss is usually acceptable. Generally, omission of tied observations decreases the power of the test.

Due to the discrete nature of S , it is generally not possible to define a rejection region that results in a test whose size is exactly equal to a prescribed α . Consequently, the approach presented for this test relies on ϕ -values rather than on defining rejection regions for the statistical conclusion. The three possible alternative hypotheses and their associated ϕ -value calculation formulas are presented in Table 10.1 on the facing page. If one assumes a normal population, the asymptotic relative efficiency (*ARE*) of the sign test relative to the t -test is $\frac{2}{\pi} \approx 0.637$. The *ARE* of the sign test relative to the t -test is also quite poor ($1/3$) for the uniform distribution (short tails). However, for the Laplace distribution (long tails), the *ARE* of the sign test in relation to the t -test is 2.

10.2.1 Confidence Interval Based on the Sign Test

A corresponding confidence interval for the median is also based on the binomial distribution by using the same assumptions as those for the one-sample sign test; however, the full sample is always used in the calculations. Assume X_1, X_2, \dots, X_n is a random sample of n observations drawn from a continuous population with an unknown median ψ . A confidence interval with a confidence level of at least $(1 - \alpha) \cdot 100\%$ can be constructed by using the k^{th} and $(n - k + 1)^{\text{st}}$ order statistics of the sample, where k is the largest value such that $\mathbb{P}(S < k) \leq \alpha/2$. For a one-sided confidence interval, k is the largest value such that $\mathbb{P}(S < k) \leq \alpha$. Order statistics are the random variables X_1, X_2, \dots, X_n rearranged in order of relative magnitude and are denoted $X_{(1)}, X_{(2)}, \dots, X_{(n)}$. That is, $X_{(1)} < X_{(2)} < \dots < X_{(n)}$. The actual confidence level is given by

$$1 - 2 \times \mathbb{P}(S < k) \text{ where } \mathbb{P}(S = x) = \binom{n}{x} \left(\frac{1}{2}\right)^n. \quad (10.1)$$

Table 10.1: Summary for testing the median — sign test

Null Hypothesis — $H_0 : \psi = \psi_0$

Test Statistic's Value — $s =$ number of observed positive differences

Alternative Hypothesis	\wp -Value Formula
$H_1 : \psi < \psi_0$	$\mathbb{P}(S \leq s H_0) = \sum_{i=0}^s \binom{n}{i} \left(\frac{1}{2}\right)^n$
$H_1 : \psi > \psi_0$	$\mathbb{P}(S \geq s H_0) = \sum_{i=s}^n \binom{n}{i} \left(\frac{1}{2}\right)^n$
$H_1 : \psi \neq \psi_0$	$\sum_{i=0}^n I(\mathbb{P}(S = i) \leq \mathbb{P}(S = s)) \cdot \binom{n}{i} \left(\frac{1}{2}\right)^n$
Recall that $I(\text{condition}) = 1$ if <i>condition</i> is true and 0 if <i>condition</i> is false.	

Clearly, k must be a positive integer since it is the subscript of an order statistic. Using (10.1) will seldom produce typical confidence levels such as 90%, 95%, or 99% exactly. Many texts provide charts to find k for the construction of confidence intervals at these typical confidence levels that are based either on always attaining a level of at least $(1 - \alpha) \times 100\%$ confidence or by providing the value of k such that the achieved confidence level is as close to $(1 - \alpha) \times 100\%$ as possible. The first approach will always return confidence intervals with a confidence level of $(1 - \alpha) \times 100\%$ or more. Roughly 50% of the confidence intervals computed with the second approach will return confidence intervals of less than the reported confidence.

The function `SIGN.test()` provided in the PASWR package returns two confidence intervals with exact confidence levels closest to the $(1 - \alpha) \times 100\%$ level specified by the user. One of these intervals has a confidence level lower than the specified level and the other has a higher confidence level than the specified level. Finally, the function uses linear interpolation between these first two intervals to give a confidence interval with the level specified by the user.

10.2.2 Normal Approximation to the Sign Test

For moderately sized samples ($n > 20$), the binomial distribution with $\pi = 0.5$ can be reasonably approximated with the normal distribution. Since $S \sim \text{Bin}(n, 0.5)$, it follows that $\mu_S = n(0.5)$ and $\sigma_S = \sqrt{n(0.5)^2}$. That is, $S \rightsquigarrow N(\mu_S, \sigma_S)$. The standardized test statistic under the assumption that $H_0 : \psi = \psi_0$ is true is

$$Z = \frac{S - n(0.5)}{\sqrt{n(0.5)^2}} \rightsquigarrow N(0, 1), \tag{10.2}$$

where S is defined as the number of positive differences among the $X_1 - \psi_0, X_2 - \psi_0, \dots, X_n - \psi_0$. See Figure 10.1 on the next page for a graph of a $\text{Bin}(20, 0.5)$ superimposed with a normal distribution with $\mu_S = 20(0.5) = 10$ and $\sigma_S = \sqrt{20(0.5)^2} = 3.16$.

The formula for calculating the observed value of the standardized test statistic as well as the three possible alternative hypotheses and their rejection regions are described in Table 10.2 on the following page.

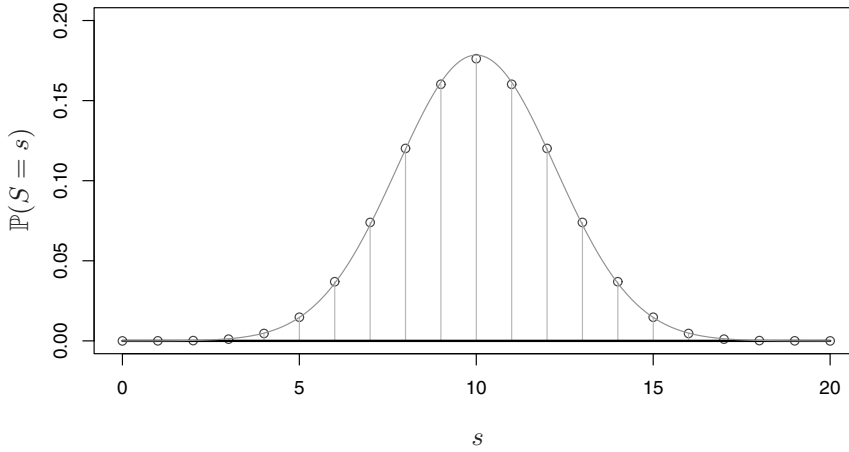


FIGURE 10.1: Graphical representation of a $Bin(20, 0.5)$ distribution and a superimposed normal distribution with $\mu_S = 20(0.5) = 10$ and $\sigma_S = \sqrt{20(0.5)^2} = 3.16$

Table 10.2: Summary for testing the median — approximation to the sign test

Null Hypothesis — $H_0 : \psi = \psi_0$

Standardized Test Statistic's Value — $z_{obs} = \frac{s \pm 0.5 - n(0.5)}{\sqrt{n(0.5)^2}}$

Alternative Hypothesis	Rejection Region
$H_1 : \psi < \psi_0$	$z_{obs} < z_\alpha$
$H_1 : \psi > \psi_0$	$z_{obs} > z_{1-\alpha}$
$H_1 : \psi \neq \psi_0$	$ z_{obs} > z_{1-\alpha/2}$
Note: The quantity ± 0.5 in the numerator of z_{obs} is the continuity correction. When $H_1 : \psi < \psi_0$, the quantity $+0.5$ is used. When $H_1 : \psi > \psi_0$, the quantity -0.5 is used. When $H_1 : \psi \neq \psi_0$, use $+0.5$ if $s < n(0.5)$ and -0.5 if $s > n(0.5)$.	

A corresponding confidence interval for the median based on (10.2) is formed with the k^{th} and $(n - k + 1)^{\text{st}}$ order statistics of the sample, where

$$k = \frac{n + 1 + z_{\alpha/2} \times \sqrt{n}}{2}. \tag{10.3}$$

For a one-sided confidence interval, replace $z_{\alpha/2}$ with z_α and solve for k . Since k is generally not an integer, it can be either rounded or truncated. To obtain a conservative estimate, one should truncate.

Example 10.1 ▷ *Sign Test: Telephone Call Times* ◁ Table 10.3 and the variable `call.time` in the data frame `Phone` contain the times in minutes of long distance telephone calls during a one month period for a small business.

Table 10.3: Long distance telephone call times in minutes (`Phone`)

i	$X_{(i)}$	i	$X_{(i)}$	i	$X_{(i)}$	i	$X_{(i)}$	i	$X_{(i)}$	i	$X_{(i)}$
1	0.2	5	0.7	9	1.3	13	2.7	17	5.6	21	9.7
2	0.2	6	0.7	10	1.7	14	4.0	18	6.1	22	9.7
3	0.2	7	0.7	11	2.1	15	4.3	19	6.7	23	12.9
4	0.2	8	0.8	12	2.1	16	5.2	20	7.0		

- Use an exact test with $\alpha = 0.05$ to see if 2.1 minutes is a representative measure of center for the telephone call lengths.
- Construct a 95% confidence interval for the population median.
- Use a normal approximation for the test used in part (a) to test if 2.1 minutes is a representative measure of center for the telephone call lengths.
- Construct a 95% confidence interval for the population median using (10.3).

Solution: First, use the function `EDA()` to assess the general shape of telephone call times. The four graphs in Figure 10.2 on the following page all lead one to the conclusion that the distribution of the long distance telephone call times is positively skewed (skewed right). Consequently, the median is a more representative measure of center than is the mean for these data.

- Use the five-step procedure to test if 2.1 minutes is a representative measure of center.

Step 1: Hypotheses — The null and alternative hypotheses to test whether or not 2.1 minutes is a representative measure of the center of telephone call times are

$$H_0 : \psi = 2.1 \text{ versus } H_1 : \psi \neq 2.1.$$

Step 2: Test Statistic — The test statistic chosen is S , where S is the number of positive differences among $X_1 - 2.1, X_2 - 2.1, \dots, X_n - 2.1$. Here, $s = 11$. Also note that since there are two instances where $x_i = \psi_0$, n is reduced from 23 to 21.

Step 3: Rejection Region Calculations — Rejection is based on the ϕ -value, so none are required.

Step 4: Statistical Conclusion — The ϕ -value is

$$\sum_{i=0}^n I(\mathbb{P}(S = i) \leq \mathbb{P}(S = s)) \cdot \binom{n}{i} \left(\frac{1}{2}\right)^n = 1.$$

```
> p.value <- sum( dbinom(0:21,21,0.5) [dbinom(0:21,21,0.5)
+ <=dbinom(11,21,0.5)] )
> p.value
[1] 1
```

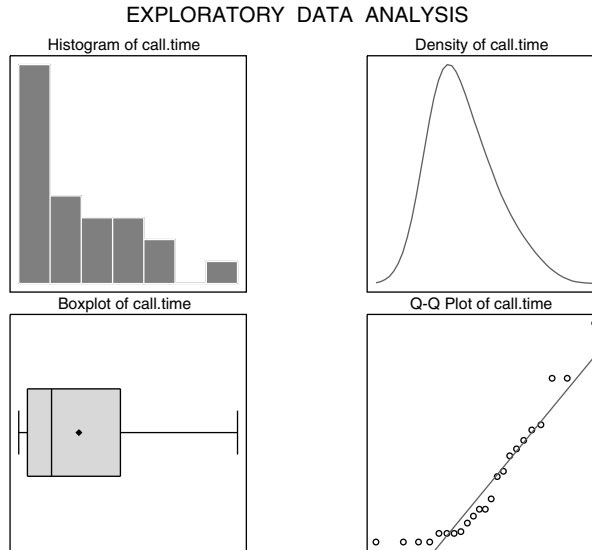


FIGURE 10.2: Graphical representation of the data in `call.time` with the function `EDA()`

Fail to reject H_0 .

Step 5: **English Conclusion** — There is not sufficient evidence to suggest the median length of long distance telephone calls is not 2.1 minutes.

(b) To construct a 95% confidence interval for the population median, start by finding the largest and smallest values of k such that $1 - 2 \times \mathbb{P}(S < k) > 0.95$ and $1 - 2 \times \mathbb{P}(S < k) < 0.95$, respectively. To find these values, use `S`, and type `1 - 2 * pbinom(0:23,23,0.5)`:

```
> round(1-2*pbinom(5:8,23,0.5),3)
[1] 0.989 0.965 0.907 0.790
```

From the `S` output, it is seen that

$$1 - 2 \times \mathbb{P}(S \leq 6) = 1 - 2 \times \mathbb{P}(S < 7) = 0.965$$

and

$$1 - 2 \times \mathbb{P}(S \leq 7) = 1 - 2 \times \mathbb{P}(S < 8) = 0.907.$$

So, use the $k = 7^{\text{th}}$ with the $n - k + 1 = 23 - 7 + 1 = 17^{\text{th}}$ order statistics to form the 96.5% confidence interval, $CI_{0.965}(\psi) = [0.7, 5.6]$, and the $k = 8^{\text{th}}$ with the $n - k + 1 = 23 - 8 + 1 = 16^{\text{th}}$ order statistics to form the 90.7% confidence interval, $CI_{0.907}(\psi) = [0.8, 5.2]$. Thus, the 95% interpolated confidence interval, $[L, U]$ is calculated such that

$$\begin{aligned} \frac{0.965 - 0.907}{0.965 - 0.95} &= \frac{0.7 - 0.8}{0.7 - L} & \frac{0.965 - 0.907}{0.965 - 0.95} &= \frac{5.6 - 5.2}{5.6 - U} \\ \Rightarrow L &= 0.73 & \Rightarrow U &= 5.49 \end{aligned}$$

So, $CI_{0.95}(\psi) = [0.73, 5.49]$.

Using the function `SIGN.test()` on the variable (`call.time`) gives the following output:

```
> attach(Phone)
> SIGN.test(call.time, md=2.1)

      One-sample Sign-Test

data:  call.time
s = 11, p-value = 1
alternative hypothesis: true median is not equal to 2.1
95 percent confidence interval:
 0.7261939 5.4952244
sample estimates:
median of x
      2.1

              Conf.Level L.E.pt U.E.pt
Lower Achieved CI      0.9069 0.8000 5.2000
Interpolated CI        0.9500 0.7262 5.4952
Upper Achieved CI      0.9653 0.7000 5.6000
```

(c) Use the five-step procedure using the normal approximation to the sign test to test if 2.1 minutes is a representative measure of center.

Step 1: **Hypotheses** — The null and alternative hypotheses to test whether or not 2.1 minutes is a representative measure of the center of telephone call times are

$$H_0 : \psi = 2.1 \text{ versus } H_1 : \psi \neq 2.1.$$

Step 2: **Test Statistic** — The test statistic chosen is S , where S is the number of positive differences among $X_1 - 2.1, X_2 - 2.1, \dots, X_n - 2.1$. Here, $s = 11$. Also note that since there are two instances where $x_i = \psi_0$, n is reduced from 23 to 21.

Step 3: **Rejection Region Calculations** — Because the standardized test statistic is distributed approximately $N(0, 1)$ and H_1 is a two-sided hypothesis, the rejection region is $|z_{obs}| \geq z_{1-0.05/2} = 1.96$. The value of the standardized test statistic is

$$z_{obs} = \frac{s \pm 0.5 - n(0.5)}{\sqrt{n(0.5)^2}} = \frac{11 - 0.5 - 21(0.5)}{\sqrt{21(0.5)^2}} = 0$$

Step 4: **Statistical Conclusion** — The ϕ -value is $2 \times \mathbb{P}(Z \geq 0) = 1$.

- I. From the rejection region, do not reject H_0 because $|z_{obs}| = 0$ is not larger than 1.96.
- II. From the ϕ -value, do not reject H_0 because the ϕ -value = 1 is larger than 0.05.

Fail to reject H_0 .

Step 5: **English Conclusion** — There is not sufficient evidence to suggest the median length of long distance telephone calls is not 2.1 minutes.

(d) To construct a confidence interval for the population median using (10.3), solve

$$\begin{aligned} k &= \frac{n + 1 + z_{\alpha/2} \times \sqrt{n}}{2} \\ &= \frac{23 + 1 - 1.96 \times \sqrt{23}}{2} = 7.3. \end{aligned}$$

By truncating k , $k = 7$. The approximate 95% confidence interval for ψ is then

$$[x_{(k)}, x_{(n-k+1)}] = [x_{(7)}, x_{(23-7+1)}] = [0.7, 5.6].$$

■

10.3 Wilcoxon Signed-Rank Test

In Section 10.2, the sign test was used to test hypotheses concerning the median. The only requirements placed on the data when using the sign test are, first, that the population from which one is sampling be continuous, and, second, that the population has a median. Since the sign test only uses the signs of the differences between each observation and the hypothesized median ψ_0 , a test that incorporates not only the signs of the differences but also their magnitudes might yield a better performance in terms of power. In fact, the test presented in this section uses both the signs of the differences between each observation and the hypothesized median as well as the magnitudes of the differences. However, in order to use this test, the **Wilcoxon signed-rank test**, one must also assume a symmetric population in addition to the assumptions for the sign test. Consequently, although the Wilcoxon signed-rank test can be used to test hypotheses concerning the median, it can be used equally well to test hypotheses regarding the mean as the assumption of symmetry in the distribution is equivalent to assuming that the mean and median are equal.

The assumption of symmetry is a rather restrictive assumption compared to those for the sign test. Nevertheless, it is certainly less restrictive than its normal analog, the t -test, which requires the data not only to come from a population that is symmetric about its median but also to come from a normal distribution. However, the Wilcoxon signed-rank test does not always perform better than the sign test. For the Laplace distribution (long-tailed), for example, the ARE of the sign test relative to the Wilcoxon signed-rank test is $4/3$. Table 10.4 lists the ARE of the sign test relative to the t -test ($ARE(S, t)$), the ARE of the Wilcoxon signed-rank test relative to the t -test ($ARE(T^+, t)$), and the ARE of the sign test relative to the Wilcoxon signed-rank test ($ARE(S, T^+)$) for the uniform distribution (short tails), normal distribution, and the Laplace distribution (long tails).

Table 10.4: Asymptotic relative efficiency comparisons

Distribution	$ARE(S, t)$	$ARE(T^+, t)$	$ARE(S, T^+)$
Uniform	1/3	1	1/3
Normal	$2/\pi \approx 0.64$	$3/\pi \approx 0.955$	2/3
Laplace	2	1.5	4/3

In summary, for large n , when testing location for symmetric populations, it is generally better to use the sign test with Laplace populations and the Wilcoxon signed-rank test for all other non-normal symmetric distributions. It can be shown that the *ARE* of the Wilcoxon signed-rank test relative to the t -test is never less than 0.864 for any continuous distribution and is ∞ for some distributions. For small n , it is not so clear which test will be better.

Given a random sample X_1, X_2, \dots, X_n taken from a continuous population that is symmetric with respect to its median ψ , under the null hypothesis $H_0 : \psi = \psi_0$, the differences, $D_i = X_i - \psi_0$, are symmetrically distributed about zero. Further, positive and negative differences of the same magnitude have the same probability of occurring. As with the sign test, if any of the d_i s are zero, they are removed from the sample before the ranks are computed, and the value of n is reduced accordingly.

To compute the Wilcoxon signed-rank statistic,

Step A: Take the absolute value of the n d_i s.

Step B: Assign the ranks to the n values from step A. If there are ties, use the **midranks**. The midrank is defined as the average rank of the tied observations.

Step C: Multiply the values in step B by the sign of the original d_i s.

Step D: Sum the positive quantities in step C. The result is denoted t^+ . The random variable (test statistic) T^+ is defined as the sum of the positive signed ranks and the random variable T^- is defined as the sum of negative signed ranks.

Provided the null hypothesis is true, $E(T^+) = E(T^-)$. When T^+ is either sufficiently small or sufficiently large, the null hypothesis is rejected. The test statistic T^+ takes on values between 0 and $n(n+1)/2$, and has a mean and variance of $n(n+1)/4$ and $n(n+1)(2n+1)/24$, respectively. The distribution of T^+ is known as the **Wilcoxon signed-rank distribution**. Although conceptually easy to understand, one needs access to extensive tables or statistical software to compute exact φ -values. Further, tabled values for T^+ are generally published only when there are no ties in the absolute values of the d_i s, $|d_i|$ for $i = 1, \dots, n$. When there are ties in the $|d_i|$ s, the S function `wilcox.test()` uses a normal approximation to compute the φ -values. It is possible to calculate exact φ -values when testing hypotheses about the median as well as to construct exact confidence intervals for the median even in the presence of ties using the function `wilcoxE.test()` from the PASWR package. The function is rather primitive and should only be used for problems with fewer than 19 observations as the memory requirements are rather large.

Example 10.2 ▷ *Trivial T^+ Distribution* ◁ What is the sampling distribution of T^+ for the trivial case where $n = 3$ and $X_1 \neq X_2 \neq X_3$?

Solution: Since there are three values ($n = 3$) that must be ranked and each d_i may have either a positive or negative sign, there are a total of $2^n = 2^3 = 8$ possible sets of signs associated with the three possible ranks (1, 2, 3). Under the null hypothesis, each of the sets of signs is equally likely to occur, and thus each has a probability of $1/8$ of occurring. Table 10.5 on the following page lists the eight possible sets of signs and Table 10.6 on the next page provides the probability distribution (**pdf**) for T^+ . ■

Table 10.5: Possible sign and rank combinations for Example 10.2 on the preceding page

	-1	+1	-1	+1	-1	+1	-1	+1
	-2	-2	+2	+2	-2	-2	+2	+2
	-3	-3	-3	-3	+3	+3	+3	+3
t^+	0	1	2	3	3	4	5	6

Table 10.6: PDF of T^+ for Example 10.2

t^+	$\mathbb{P}(T^+ = t^+)$
0	1/8
1	1/8
2	1/8
3	2/8
4	1/8
5	1/8
6	1/8

R can compute quantiles (`qsignrank()`), the density function (`dsignrank()`), the distribution function (`psignrank()`), and random numbers (`rsignrank()`) from the Wilcoxon signed-rank distribution. For example, the probabilities in Table 10.6 can be generated with `dsignrank(0:6,3)`. To obtain further help, type `?dsignrank` at the R prompt. S-PLUS has the function `psignrank()`, but it was not documented at the time of writing. Also, S-PLUS did not have the functions `dsignrank()`, `qsignrank()`, nor `rsignrank()`.

Due to the discrete nature of T^+ , it is generally not possible to define a rejection region that results in a test whose size is exactly equal to the prescribed α . Consequently, the approach presented for this test relies on φ -values rather than on defining rejection regions for the statistical conclusion. The three possible alternative hypotheses and their associated φ -value calculation formulas are presented in Table 10.7 on page 415. The φ -value formulas given in Table 10.7 can be used to calculate exact φ -values with S's `psignrank()` function when there are no ties among the non-zero $|d_i|$ s. In the presence of ties, the S function `wilcox.test()` uses Table 10.9 on page 419 with a correction factor. The formulas in Table 10.7 on page 415 are still valid when there are ties in the non-zero $|d_i|$ s; however, the exact conditional distribution of T^+ when ties are present is not a base function of S. Example 10.3 shows how S can be used to compute the exact φ -value for the conditional distribution of T^+ (the distribution of T^+ with ties in the non-zero $|d_i|$ s).

Example 10.3 ▷ *Wilcoxon Signed-Rank Test: Pool pH* ◁ A lifeguard is told to maintain the pH of a 50 m pool at 7.25. He takes pH measurements at each of the four corners of the pool and gets 7.2, 7.3, 7.3, and 7.4. Calculate the φ -value for testing the hypothesis that the median pH is greater than 7.25 using the exact conditional distribution for T^+ .

Solution: If the data are symmetric, a Wilcoxon signed-rank test may be appropriate. A visual inspection of the pH measurements reveals they are symmetric around 7.3. The creation of a density plot to verify this assumption is left to the reader. The steps for

carrying out the Wilcoxon signed-rank test are to, first, create the d_i values that equal $x_i - \psi_0$ for $i = 1, \dots, n$. Next, take the absolute value of the n d_i s and assign the ranks to the n values. If there are ties, use the **midranks**. Then, multiply the values of the ranks of the $|d_i|$ s by the sign of the original d_i s. Finally, sum the resulting positive quantities to obtain t^+ .

```
> PH <- c(7.2,7.3,7.3,7.4)           # Enter data
> DIFF <- PH-7.25                    # Create differences (DIFF)
> absD <- abs(DIFF)                  # Absolute value of DIFF (absD)
> rankabsD <- rank(absD)             # Rank the absD values
> signD <- sign(DIFF)                # Store the signs of DIFF
> signrank <- rankabsD*signD         # Create a vector of signed ranks
> tp <- sum(signrank[signrank>0])    # Calculate t+
> tp
[1] 8
```

After t^+ is calculated, the distribution of T^+ must be enumerated to find the p -value.

```
> n <- length(DIFF)
> signs <- as.matrix(expand.grid(rep(list(0:1), n)))
> signs                               # 1s represent positive ranks
  Var1 Var2 Var3 Var4
[1,]   0   0   0   0
[2,]   1   0   0   0
[3,]   0   1   0   0
[4,]   1   1   0   0
[5,]   0   0   1   0
[6,]   1   0   1   0
[7,]   0   1   1   0
[8,]   1   1   1   0
[9,]   0   0   0   1
[10,]  1   0   0   1
[11,]  0   1   0   1
[12,]  1   1   0   1
[13,]  0   0   1   1
[14,]  1   0   1   1
[15,]  0   1   1   1
[16,]  1   1   1   1

> mat <- matrix(rankabsD)           # Put rankabsD in matrix form
> mat
  [,1]
[1,]  2
[2,]  2
[3,]  2
[4,]  4
```

After the matrix listing the locations of the positive ranks with 1s and the locations of negative ranks with 0s is created (**signs**), matrix multiplication is used to sum the positive ranks to get the distribution of T^+ , where **mat** contains the ranks of the absolute values of the d_i s:


```

> Tp <- signs%**mat           # (16X4)*(4X1) = 16X1 vector of T+
> Tp <- sort(Tp)              # Sort the distribution of T+
> SampDist <- table(Tp)/2^n
> SampDist                    # Sampling distribution of T+
Tp
  0      2      4      6      8     10
0.0625 0.1875 0.2500 0.2500 0.1875 0.0625

```

Since H_1 is an upper one-sided hypothesis, the ϕ -value is the sum of the values of the distribution of T^+ that are greater than or equal to the value of our test statistic t^+ . In this case, the t^+ was 8, so the ϕ -value is

$$\phi\text{-value} = \mathbb{P}(T^+ = 8) + \mathbb{P}(T^+ = 10) = 0.1875 + 0.0625 = 0.25.$$

```

> p.value <- sum(Tp>=tp)/2^n   # Calculate p-value
> p.value
[1] 0.25

```

This ϕ -value can also be found using the function `wilcoxE.test()` from the PASWR package. Note that the function `wilcox.test()` cannot be used because it cannot compute exact ϕ -values when there are ties in the data.

```

> wilcoxE.test(PH, mu=7.25, alternative="greater")

```

Wilcoxon Signed Rank Test

```

data: PH
t+ = 8, p-value = 0.25
alternative hypothesis: true median is greater than 7.25
93.75 percent confidence interval:
 7.25 Inf
sample estimates:
(pseudo)median
      7.3

```

The Wilcoxon signed-rank test may also be used for testing whether the median difference (ψ_0) between two dependent populations (X and Y) is equal to some value, $H_0 : \psi_D = \psi_0$. For dependent samples, $D_i = X_i - Y_i - \psi_0$ instead of $D_i = X_i - \psi_0$. The computation of T^+ for dependent samples follows the same steps as those for a single sample.

10.3.1 Confidence Interval for ψ Based on the Wilcoxon Signed-Rank Test

Since $\{X_1, X_2, \dots, X_n\}$ are random variables from a symmetric distribution with median ψ , the pairwise averages $\bar{x}_{ij} = \frac{x_i + x_j}{2}$, where $1 \leq i \leq j \leq n$, are also symmetrically distributed about the median ψ . There are a total of $n(n+1)/2$ of these \bar{x}_{ij} s, frequently called the **Walsh averages**. The $n(n+1)/2$ Walsh averages can be split into $\binom{n}{2}$ means, \bar{x}_{ij} , where $i \neq j$ and n means, \bar{x}_{ii} for $i = 1, \dots, n$. When the Walsh averages are ordered from smallest to largest, the k^{th} and $(\frac{n(n+1)}{2} - k + 1)^{\text{st}}$ order statistics are the lower and upper endpoints of a confidence interval with a confidence level of at least $(1 - \alpha) \cdot 100\%$, where k is the largest value such that $\mathbb{P}(T^+ < k) \leq \alpha/2$. For a one-sided confidence interval, k is the largest value such that $\mathbb{P}(T^+ < k) \leq \alpha$. Again, k is a positive integer since it is the

Table 10.7: Summary for testing the median — Wilcoxon signed-rank test
 Null Hypothesis — $H_0 : \psi = \psi_0$

Test Statistic's Value — t^+ = sum of the positive ranked differences

Alternative Hypothesis	ϕ -Value Formula
$H_1 : \psi < \psi_0$	$\mathbb{P}(T^+ \leq t^+ H_0)$
$H_1 : \psi > \psi_0$	$\mathbb{P}(T^+ \geq t^+ H_0) = 1 - \mathbb{P}(T^+ \leq t^+ - 1 H_0)$
$H_1 : \psi \neq \psi_0$	$2 \times \min \{ \mathbb{P}(T^+ \leq t^+), 1 - \mathbb{P}(T^+ \leq t^+ - 1), 0.5 \}$

subscript of an order statistic. The exact confidence level is $1 - 2\mathbb{P}(T^+ < k)$ for a two-sided confidence interval and $1 - \mathbb{P}(T^+ < k)$ for a one-sided confidence interval.

When there are no d_i s ($x_i - \psi_0$) that equal zero, as well as no x_i s that equal zero, testing $H_0 : \psi = \psi_0$ with the procedures described in Section 10.3 yields an equivalent acceptance region to that produced by the confidence interval based on the Walsh averages. If this is not the case, the regions are no longer equivalent.

For the dependent case, use the $n(n + 1)/2$ dependent Walsh averages $\overline{x - y_{ij}} = [(x_i - y_i) + (x_j - y_j)]/2$. In this case, $d_i = x_i - y_i - \psi_0$. Here the equivalence between the acceptance region of the hypothesis test and the confidence interval created based on the Walsh averages exists only when $d_i \neq 0$ and $x_i - y_i \neq 0, i = 1, \dots, n$.

Example 10.4 ▷ *Wilcoxon Signed-Rank Test: Waiting Times* ◁ A statistician records how long he must wait for his bus each morning. This information is recorded in Table 10.8 on the following page and in the data frame `Wait`.

- (a) Test to see if his median waiting time is less than 6 minutes.
- (b) Compute an upper 95% confidence interval for the median, ψ .

Solution: Before using the Wilcoxon signed-rank test, a quick check on the assumption of symmetry is made with a boxplot in Figure 10.3 on the next page. Since the boxplot does appear symmetric, it is legitimate to proceed with a Wilcoxon signed-rank test.

- (a) Use the five-step procedure to test if the median waiting time is less than 6 minutes.

Step 1: **Hypotheses** — The null and alternative hypotheses to test if the median waiting time is less than 6 minutes are

$$H_0 : \psi = 6 \text{ versus } H_1 : \psi < 6.$$

Step 2: **Test Statistic** — The test statistic chosen is T^+ , where T^+ is the Wilcoxon signed-rank statistic. Here, the observed value of T^+ is $t^+ = 28$.

Step 3: **Rejection Region Calculations** — Rejection is based on the ϕ -value, so none are required.

Table 10.8: Waiting times in minutes (**Wait**)

x_i	$d_i = x_i - 6$	$ d_i $	$\text{sign}(d_i)$	rank $ d_i $	signed ranks
8.0	2.0	2.0	+	6	6
2.1	-3.9	3.9	-	12	-12
3.8	-2.2	2.2	-	7	-7
8.6	2.6	2.6	+	8	8
7.3	1.3	1.3	+	4	4
6.1	0.1	0.1	+	1	1
1.4	-4.6	4.6	-	13	-13
2.9	-3.1	3.1	-	10	-10
5.5	-0.5	0.5	-	2	-2
2.7	-3.3	3.3	-	11	-11
4.8	-1.2	1.2	-	3	-3
4.6	-1.4	1.4	-	5	-5
1.0	-5.0	5.0	-	14	-14
8.7	2.7	2.7	+	9	9
0.8	-5.2	5.2	-	15	-15

$$t^+ = \mathbf{28}$$

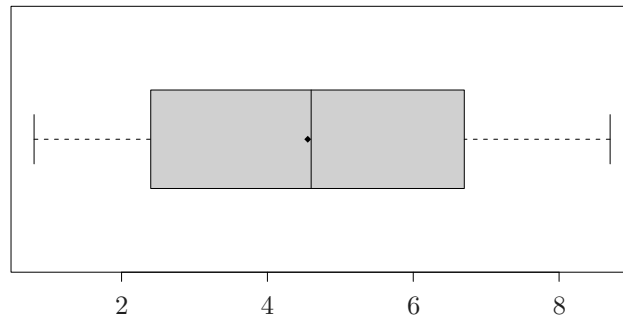


FIGURE 10.3: Horizontal boxplot of bus waiting times in minutes

Step 4: **Statistical Conclusion** — The φ -value is $\mathbb{P}(T^+ \leq 28) = 0.03649902$, which can be obtained by typing `psignrank(28,15)`.

Reject H_0 .

Step 5: **English Conclusion** — There is sufficient evidence to suggest the median waiting time is less than 6 minutes.

S performs this test with the function `wilcox.test()`:

```
> attach(Wait)
> wilcox.test(wt, mu=6, alternative="less")

      Wilcoxon signed rank test

data:  wt
V = 28, p-value = 0.0365
alternative hypothesis: true location is less than 6
```

Note that S denotes the statistic t^+ with a V .

(b) To compute an upper 95% confidence interval for ψ , first determine the $n(n+1)/2 = 15(15+1)/2 = 120$ Walsh averages. To ease the drudgery of 120 calculations of means, use S:

```
> n2means <- apply(SRS(wt,2),1, mean) #Computing the n choose 2 means
> WalshAverages <- c(wt, n2means)
```

Next, find the largest value k such that $\mathbb{P}(T^+ < k) \leq 0.05$. This can be accomplished in two ways:

(1) Type `qsignrank(α , n)` into R:

```
> qsignrank(0.05,15)
[1] 31
```

(2) Visually inspect `psignrank(0:n*(n+1)/2, n)` for the largest value k such that $\mathbb{P}(T^+ < k) \leq \alpha$. Note that the first pair $(k-1, \mathbb{P}(T^+ < k))$ of the output shown is $(28, 0.036)$, and the pair that gives the answer is $(30, 0.047)$, which implies $k-1 = 30$ or $k = 31$.

```
> psi <- round(psignrank(28:33,15),3)
> names(psi) <- 28:33
> psi
   28   29   30   31   32   33
0.036 0.042 0.047 0.053 0.060 0.068
```

Either (1) or (2) can be used with R while (2) must be used with S-PLUS. Note that if method (1) is used, that the exact confidence level will be $1 - \text{psignrank}(30, 15)$ for an upper one-sided confidence interval. The 95.27% confidence interval where $k = 31$ is then

$$\left(-\infty, \bar{x}_{\left(\frac{n(n+1)}{2} - k + 1\right)}\right] = \left(-\infty, \bar{x}_{(90)}\right] = \left(-\infty, 5.8\right].$$

```
> SWA <- sort(WalshAverages)
> SWA[90]
[1] 5.8
```

This may be done directly with the argument `conf.int=TRUE` in the `wilcox.test()` function if one is using R:

```
> wilcox.test(wt, mu=6, alternative="less", conf.int=TRUE)
```

```
Wilcoxon signed rank test
```

```
data: wt
V = 28, p-value = 0.0365
alternative hypothesis: true location is less than 6
95 percent confidence interval:
 -Inf 5.8
sample estimates:
(pseudo)median
      4.625
```

The answer can also be computed with the function `wilcoxE.test` from the PASWR package:

```
> wilcoxE.test(wt, mu=6, alternative="less")
```

```
Wilcoxon Signed Rank Test
```

```
data: wt
t+ = 28, p-value = 0.0365
alternative hypothesis: true median is less than 6
95.26978 percent confidence interval:
 -Inf 5.8
sample estimates:
(pseudo)median
      4.625
```



10.3.2 Normal Approximation to the Wilcoxon Signed-Rank Test

For moderately sized samples ($n > 15$), the sampling distribution of T^+ can be reasonably approximated with the normal distribution that has a mean and standard deviation $n(n+1)/4$ and $\sqrt{n(n+1)(2n+1)/24}$, respectively. That is, $T^+ \sim N(n(n+1)/4, \sqrt{n(n+1)(2n+1)/24})$. The standardized test statistic under the assumption that $H_0 : \psi = \psi_0$ is true is

$$Z = \frac{T^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \sim N(0, 1). \quad (10.4)$$

See Figure 10.4 on the facing page for a graph of the Wilcoxon signed-rank distribution for $n = 15$ superimposed by a normal distribution with $\mu = n(n+1)/4 = 60$ and $\sigma = \sqrt{n(n+1)(2n+1)/24} = 17.61$.

The formula for calculating the observed value of the standardized test statistic as well as the three possible alternative hypotheses and their rejection regions are described in Table 10.9 on the next page. If there are ties in the $|d_i|$ s, the variance of T^+ is reduced to

$$\frac{n(n+1)(2n+1)}{24} - \frac{\sum_{j=1}^g t_j(t_j-1)(t_j+1)}{48} \quad (10.5)$$

where g denotes the number of tied groups of non-zero $|d_i|$ s and t_j is the size of tied group j . In (10.5), an untied observation is considered to be a tied group of size one. In the event that no ties exist, $g = n$ and $t_j = 1$ for $j = 1, \dots, n$, which produces a correction factor of zero.

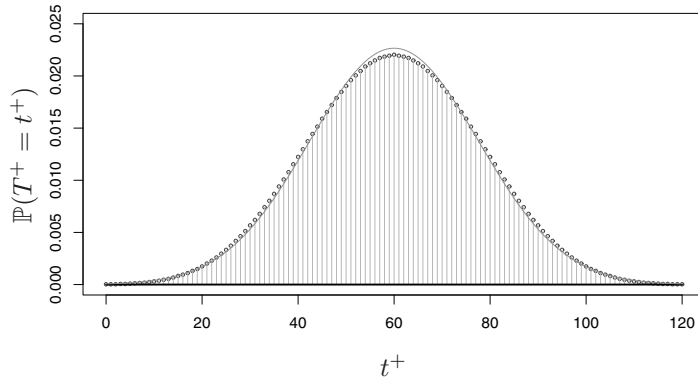


FIGURE 10.4: Graphical representation of the Wilcoxon signed-rank distribution for $n = 15$ superimposed by a normal distribution with $\mu = n(n + 1)/4 = 60$ and $\sigma = \sqrt{n(n + 1)(2n + 1)/24} = 17.61$

Table 10.9: Summary for testing the median — normal approximation to the Wilcoxon signed-rank test

Null Hypothesis — $H_0 : \psi = \psi_0$

Standardized Test Statistic's Value — $z_{obs} = \frac{t^+ \pm 0.5 - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24 - CF}}$

Correction Factor — $CF = \sum_{j=1}^g t_j(t_j - 1)(t_j + 1)/48$

Alternative Hypothesis	Rejection Region
$H_1 : \psi < \psi_0$	$z_{obs} < z_\alpha$
$H_1 : \psi > \psi_0$	$z_{obs} > z_{1-\alpha}$
$H_1 : \psi \neq \psi_0$	$ z_{obs} > z_{1-\alpha/2}$

Note: The quantity ± 0.5 in the numerator of z_{obs} is the continuity correction. When $H_1 : \psi < \psi_0$, the quantity $+0.5$ is used. When $H_1 : \psi > \psi_0$, the quantity -0.5 is used. When $H_1 : \psi \neq \psi_0$, use $+0.5$ if $t^+ < n(n + 1)/4$ and -0.5 if $t^+ > n(n + 1)/4$.

A corresponding two-sided confidence interval for the median based on (10.4) are the k^{th} and $(n(n + 1)/2 - k + 1)^{\text{st}}$ ordered Walsh averages, where

$$k = 0.5 + \frac{n(n + 1)}{4} + z_{\alpha/2} \sqrt{\frac{n(n + 1)(2n + 1)}{24}}. \tag{10.6}$$

For a one-sided confidence interval, replace $z_{\alpha/2}$ with z_α . Since k is generally not an integer, it can be either rounded or truncated. To obtain a conservative estimate, one should truncate.

Example 10.5 ▷ *Wilcoxon Signed-Rank Test: TV Effects* ◁ Gibbons (1997) provides the following data regarding aggressive behavior in relation to exposure to violent television programs with the following exposition:

... a group of children are matched as well as possible as regards home environment, genetic factors, intelligence, parental attitudes, and so forth, in an effort to minimize factors other than TV that might influence a tendency for aggressive behavior. In each of the resulting 16 pairs, one child is randomly selected to view the most violent shows on TV, while the other watches cartoons, situation comedies, and the like. The children are then subjected to a series of tests designed to produce an ordinal measure of their aggression factors. (pages 143–144)

The data that were collected are presented in Table 10.10 on the facing page and stored in data frame `Aggression`, where x_i represents aggression test scores for the children who watched violent programming (`violence`) and y_i represents aggression test scores for the children who watched non-violent television programs (`noviolence`).

- Confirm that the distribution is symmetric.
- Test whether the median difference for aggression test scores for pairs of children is greater than zero using a significance level of $\alpha = 0.05$ with the normal approximation to the Wilcoxon signed-rank test.
- Use the function `wilcoxE.test()` to report the exact p -value and the lower one-sided confidence interval for the hypothesis in (b).
- Construct a lower one-sided confidence interval with confidence level of at least 95% using the normal approximation to find k .

Solution: The answers are as follows:

- Before using the Wilcoxon signed-rank test, a quick check on the assumption of symmetry is made with a boxplot in Figure 10.5. Since the boxplot does appear symmetric, it is legitimate to proceed with a Wilcoxon signed-rank test.

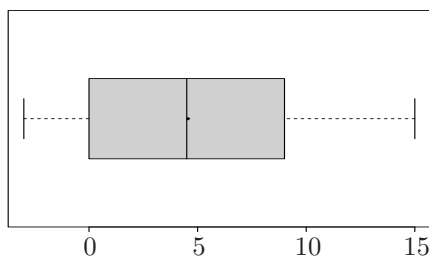


FIGURE 10.5: Horizontal boxplot of differences of aggression scores

- Use the five-step procedure to test if the median difference for aggression scores for pairs of children is greater than zero.

Table 10.10: Aggression test scores (**Aggression**)

Pair	x_i	y_i	$d_i = x_i - y_i$	$ d_i $	$\text{sign}(d_i)$	$\text{rank} d_i $	signed ranks
1	35	26	9	9	+	12.5	12.5
2	30	28	2	2	+	4.5	4.5
3	15	16	-1	1	-	1.5	-1.5
4	20	16	4	4	+	8	8
5	25	16	9	9	+	12.5	12.5
6	14	16	-2	2	-	4.5	-4.5
7	37	32	5	5	+	9	9
8	26	24	2	2	+	4.5	4.5
9	36	30	6	6	+	10	10
10	40	33	7	7	+	11	11
11	35	20	15	15	+	16	16
12	20	19	1	1	+	1.5	1.5
13	16	19	-3	3	-	7	-7
14	21	10	11	11	+	15	15
15	17	7	10	10	+	14	14
16	15	17	-2	2	-	4.5	-4.5

$$t^+ = \mathbf{118.5}$$

Step 1: **Hypotheses** — The null and alternative hypotheses to test if the median difference for aggression scores for pairs of children is greater than zero are

$$H_0 : \psi_D = 0 \text{ versus } H_1 : \psi_D > 0.$$

Step 2: **Test Statistic** — The test statistic chosen is T^+ , where T^+ is the Wilcoxon signed-rank statistic. Here, the observed value of T^+ is $t^+ = 118.5$.

Step 3: **Rejection Region Calculations** — Because the standardized test statistic is distributed approximately $N(0, 1)$ and H_1 is an upper one-sided hypothesis, the rejection region is $z_{obs} > z_{1-0.05} = 1.645$. Because there are three groups of ties ($g = 3$) where the sizes of the tied groups are 2, 4, and 2, the correction factor is

$$\begin{aligned} CF &= \sum_{j=1}^3 t_j(t_j - 1)(t_j + 1)/48 \\ &= [2(2 - 1)(2 + 1) + 4(4 - 1)(4 + 1) + 2(2 - 1)(2 + 1)] / 48 \\ &= [6 + 60 + 6] / 48 \\ &= 72 / 48 = 3/2 \end{aligned}$$

The value of the standardized test statistic is

$$\begin{aligned} z_{obs} &= \frac{t^+ \pm 0.5 - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24 - CF}} \\ &= \frac{118.5 - 0.5 - 16(16+1)/4}{\sqrt{16(16+1)(2(16)+1)/24 - 3/2}} \\ &= 2.591 \end{aligned}$$

Step 4: **Statistical Conclusion** — The ϕ -value is $\mathbb{P}(Z \geq 2.591) = 0.0048$

- I. From the rejection region, reject H_0 because $z_{obs} = 2.591$ is more than 1.645.
- II. From the ϕ -value, reject H_0 because the ϕ -value = 0.0048 is less than 0.05.

Reject H_0 .

Step 5: **English Conclusion** — There is sufficient evidence to suggest that children who view violent television programs have higher aggression test scores than children who view non-violent television programs.

S commands to compute the test follow. Note that S automatically uses a normal approximation to the distribution of T^+ when there are ties in the $|d_i|$ s as well as a correction factor for the variance of T^+ .

```
> attach(Aggression)
> wilcox.test(violence, noviolence, paired=TRUE, alternative="greater")
```

Wilcoxon signed rank test with continuity correction

```
data: violence and noviolence
V = 118.5, p-value = 0.00479
alternative hypothesis: true location shift is greater than 0
```

Warning message:

```
In wilcox.test.default(violence, noviolence, paired = TRUE,
  alternative = "greater") : cannot compute exact p-value with ties
```

(c) From the output of `wilcoxE.exact()`, the ϕ -value is 0.003265 and the lower 95.21% confidence interval is $[2, \infty)$:

```
> wilcoxE.test(violence, noviolence, paired=TRUE, alternative="greater")
```

Wilcoxon Signed Rank Test (Dependent Samples)

```
data: violence and noviolence
t+ = 118.5, p-value = 0.003265
alternative hypothesis: true median difference is greater than 0
95.20569 percent confidence interval:
```

2 Inf

sample estimates:

(pseudo)median

4.5

(d) The paired differences are stored in PD and the sorted Walsh averages are in SWA. Using (10.6), k is calculated to be 36, and the k^{th} Walsh average is determined to be 2. Therefore, the 95% confidence interval for ψ_D is $[2, \infty)$.

```
> PD <- violence-noviolence
> n2means <- apply(SRS(PD,2),1, mean) # Computing the n choose 2 means
> SWA <- sort(c(PD, n2means))        # Sorted Walsh averages
> n <- length(PD)
> k <- 0.5 + n*(n+1)/4 + qnorm(.05)*sqrt(n*(n+1)*(2*n+1)/24)
> k <- floor(k)
> k
[1] 36
> SWA[k]                            # Walsh average k
[1] 2
```

Another way to achieve the same result is with the function `outer()`, which applies the third argument (" $+$ ") to the first two vectors in an element-wise manner to create an array, and `!lower.tri`, which returns the values of the upper triangular matrix containing double the Walsh averages. Finally, the upper triangular matrix is divided by two, and then sorted to calculate the values for the sorted Walsh averages.

```
> ADD <- outer(PD, PD, "+")
> SWA2 <- sort(ADD[!lower.tri(ADD)])/2
> SWA2[k]                            # Walsh average k
[1] 2
```



10.4 The Wilcoxon Rank-Sum or the Mann-Whitney U -Test

The **Wilcoxon rank-sum** test is due to Wilcoxon (1945). Its widespread use is due in large part to Mann and Whitney, who proposed another test, the **Mann-Whitney U -test**, which is equivalent to the Wilcoxon rank-sum test. However, be aware that many combinations of names with either some or all of Mann, Whitney, and Wilcoxon are all typically referring to some variation of the same test. The two-sample Wilcoxon rank-sum test assumes that data come from two independent random samples X_1, X_2, \dots, X_n and Y_1, Y_2, \dots, Y_m of sizes n and m , respectively, where the underlying distributions of X_1, X_2, \dots, X_n and Y_1, Y_2, \dots, Y_m have the same shape. Note that the assumption of identical underlying shapes implies that the variances are also equal. No further assumptions other than continuous data, which is at least on an ordinal scale, are made with the two-sample Wilcoxon rank-sum test. Because the underlying distributions of X and Y are assumed to be identical in the null hypothesis, this test can apply to means, medians, or any other quantile.

If two random samples of size n and m are drawn from two identical populations, all $N = n + m$ observations can be regarded as a single sample from some common population. Further, if the N observations are ordered in a single sequence according to relative magnitude, one expects the X s and Y s to be well mixed in the ordered sequence that represents the sample data. That is, an arrangement of the data where most of the X s are smaller than the Y s, or vice versa, would suggest two distinct populations and not one common population. The Wilcoxon rank-sum statistic, W , is computed by

1. Forming a single sample of all $n + m$ observations.
2. Assigning ranks to the combined sample.
3. Summing the X ranks in the combined sample.

Provided the null hypothesis of identical populations is true, typically denoted $H_0 : F_X(x) = F_Y(x)$ for all x , all $\binom{N}{n}$ assignments of the X ranks are equally likely, each having probability $1/\binom{N}{n}$. Consequently, W values that are either too small or too large will cause the null hypothesis to be rejected. W takes on integer values ranging from $n(n+1)/2$ to $n(2N-n+1)/2$ when no ties are present in the ranks. The sampling distribution of W , known as the **Wilcoxon rank-sum** distribution, is symmetric about its mean value $n(N+1)/2$ and has a variance of $nm(N+1)/12$.

Due to the discrete nature of W , it is generally not possible to define a rejection region that results in a test whose size is exactly equal to the prescribed α . Consequently, the approach presented for this test relies on φ -values rather than on defining rejection regions for the statistical conclusion. The three possible alternative hypotheses and their associated φ -value calculation formulas are presented in Table 10.11.

Table 10.11: Summary for testing equality of medians — Wilcoxon rank-sum test

$$\text{Null Hypothesis — } H_0 : \psi_X - \psi_Y = \delta_0$$

Test Statistic's Value — $w =$ sum of the ranked x s in the combined sample

Alternative Hypothesis	φ -Value Formula
$H_1 : \psi_X - \psi_Y = \delta_0$	$\mathbb{P}(W \leq w H_0)$
$H_1 : \psi_X - \psi_Y > \delta_0$	$\mathbb{P}(W \geq w H_0) = 1 - \mathbb{P}(W \leq w - 1 H_0)$
$H_1 : \psi_X - \psi_Y \neq \delta_0$	$2 \times \min \{ \mathbb{P}(W \leq w), 1 - \mathbb{P}(W \leq w - 1), 0.5 \}$

A closely related statistic to W proposed by Mann and Whitney, typically denoted by U , is defined as the total number of times the pair (x_i, y_j) contains an x value greater than the y value for all (i, j) . The relationship between W and U can be expressed as $U = W - n(n+1)/2$, and generalized to include tied ranks by assigning $1/2$ to all ties.

The φ -value formulas given in Table 10.11 can be used to calculate exact φ -values with S's `pwilcox()` function when there are no ties among the ranks. However, care needs to be taken as R and S-PLUS use different definitions for the Wilcoxon rank-sum distribution. The S-PLUS definition of the Wilcoxon rank-sum distribution corresponds to the distribution of W , while the R definition of the Wilcoxon rank-sum distribution corresponds to the distribution of U . In the presence of ties, the S-PLUS function `wilcox.test()` uses Table 10.13 on page 431 with a correction factor, while the R function `wilcox.test()` uses Table 10.14 on page 431 with the same correction factor. The formulas in Table 10.11 are still valid when there are ties in the ranks; however, the exact conditional distribution of W when ties are present is not readily available in S. Example 10.7 on page 426 shows how S can be used to compute the exact φ -value for the conditional distribution of W (the distribution of W with ties in the ranks).

Table 10.12: Summary for testing equality of medians — Mann-Whitney U -test

Null Hypothesis — $H_0 : \psi_X - \psi_Y = \delta_0$

Test Statistic's Value — $u =$ number of times x precedes y
in the pairs (x_i, y_j) for all (i, j)

Alternative Hypothesis	\wp -Value Formula
$H_1 : \psi_X - \psi_Y = \delta_0$	$\mathbb{P}(U \leq u H_0)$
$H_1 : \psi_X - \psi_Y > \delta_0$	$\mathbb{P}(U \geq u H_0) = 1 - \mathbb{P}(U \leq u - 1 H_0)$
$H_1 : \psi_X - \psi_Y \neq \delta_0$	$2 \times \min\{\mathbb{P}(U \leq u), 1 - \mathbb{P}(U \leq u - 1), 0.5\}$

Example 10.6 ▷ *W and U Sampling Distributions* ◁ Assume the values $\mathbf{x} = \{2, 5\}$ and $\mathbf{y} = \{9, 12, 14\}$ are two independent random samples from independent distributions that are assumed to be equal in shape. Enumerate the sampling distributions of W and U .

Solution: Start by reading the values of \mathbf{x} and \mathbf{y} into vectors labeled \mathbf{x} and \mathbf{y} , respectively:

```
> x <- c(2,5)
> y <- c(9,12,14)
> n <- length(x)
> m <- length(y)
> N <- n + m
> r <- rank(c(x, y))
> u <- sum(r[seq(along = x)]) - n*(n + 1)/2      # observed u value
> w <- sum(r[seq(along = x)])                  # observed w value
> val <- SRS(r, n)                            # possible rankings for X
> W <- apply(val,1, sum)                       # W values
> U <- W - n*(n + 1)/2                        # U values
> display <- cbind(val, W, U)                 # X rankings with W and U
> display
```

	W	U
[1,]	1 2 3 0	
[2,]	1 3 4 1	
[3,]	2 3 5 2	
[4,]	1 4 5 2	
[5,]	2 4 6 3	
[6,]	3 4 7 4	
[7,]	1 5 6 3	
[8,]	2 5 7 4	
[9,]	3 5 8 5	
[10,]	4 5 9 6	

Note that the values of W are between $n(n + 1)/2 = 2(2 + 1)/2 = 3$ and $n(2N - n + 1)/2 = 2[(2)(5) - 2 + 1]/2 = 9$.

```
> table(W)/choose(5,2)      # Sampling distribution of W
W
 3  4  5  6  7  8  9
0.1 0.1 0.2 0.2 0.2 0.1 0.1

> dwilcox(3:9,2,3)          # Produces distribution of W in S-PLUS
[1] 0.1 0.1 0.2 0.2 0.2 0.1 0.1
```

Note that the values of U are between 0 and $n \cdot m = 6$.

```
> table(U)/choose(5,2)      # Sampling distribution of U
U
 0  1  2  3  4  5  6
0.1 0.1 0.2 0.2 0.2 0.1 0.1

> dwilcox(0:6,2,3)          # Produces distribution of U in R
[1] 0.1 0.1 0.2 0.2 0.2 0.1 0.1
```

Example 10.7 ▷ *Wilcoxon Rank-Sum φ -Value: Pool pH* ◁ Lifeguards are told to maintain the pH of a 50 m pool at 7.25. The pool manager takes pH measurements at each of the four corners of the pool before the pool opens on two consecutive days. Calculate the φ -value for testing the hypothesis that the difference in median pH readings is zero using the exact conditional distribution of W . (The same underlying distribution assumption can be verified graphically.) The pH readings are Day 1 (x): {7.2, 7.2, 7.3, 7.3} and Day 2 (y): {7.3, 7.3, 7.4, 7.4}.

Solution: Use S to calculate the φ -values:

```
> x <- c(7.2,7.2,7.3,7.3)
> y <- c(7.3,7.3,7.4,7.4)
> n <- length(x)
> m <- length(y)
> N <- n + m
> r <- rank(c(x, y))
> w <- sum(r[seq(along = x)])      # observed w value
> w
[1] 12
> val <- SRS(r, n)                # possible rankings
> W <- apply(SRS(r, n),1, sum)    # W values

> table(W)/choose(8,4)
W
      12      15      18      21      24
0.08571429 0.22857143 0.37142857 0.22857143 0.08571429
```

Since H_1 is a two-sided hypothesis, the φ -value is

$$2 \times \min\{\mathbb{P}(W \leq w), 1 - \mathbb{P}(W \leq w - 1), 0.5\}.$$

In this case, w was 12, so $\mathbb{P}(W \leq 12) = 0.0857$ and $1 - \mathbb{P}(W \leq 11) = 1$. It follows that the φ -value is $2 \times 0.0857 = 0.1714$.

```
> p.value <- 2*(sum(W <= w)/choose(N, n))
> p.value
[1] 0.1714286
```

The output from the `wilcoxE.test()` function is

```
> wilcoxE.test(x, y)
```

```
Wilcoxon Rank Sum Test
```

```
data: x and y
```

```
w = 12, p-value = 0.1714
```

```
alternative hypothesis: true median is not equal to 0
```

```
82.85714 percent confidence interval:
```

```
-0.2 0.0
```

```
sample estimates:
```

```
difference in location
```

```
-0.1
```



10.4.1 Confidence Interval Based on the Mann-Whitney U -Test

A confidence interval with a confidence level of at least $(1 - \alpha) \cdot 100\%$ for the shift Δ from population X , from which a sample $x_i, i = 1, \dots, n$ is taken, to population Y , from which a sample $y_j, j = 1, \dots, m$ is taken, can be constructed by using the k^{th} and $(nm - k + 1)^{\text{st}}$ order statistics from the nm differences $x_i - y_j$ where k is the largest value such that $\mathbb{P}(U < k) \leq \alpha/2$. For a one-sided confidence interval, k is the largest value such that $\mathbb{P}(U < k) \leq \alpha$. The exact confidence level is given by $1 - 2 \times \mathbb{P}(U < k)$ for a two-sided confidence interval and $1 - \mathbb{P}(U < k)$ for a one-sided confidence interval. Clearly, k must be a positive integer since it is the subscript of an order statistic.

Example 10.8 ▷ *Confidence Interval for Difference in Medians* ◁ Eight spring piglets are randomly assigned to two different groups and are fed two different diets (A and B). After four weeks, the weight gains in pounds for the piglets eating each diet are recorded. Find a 90% confidence interval for the median difference in weight gains for the piglets eating each diet. Be sure to verify the assumption of identical underlying distributions except for a shift that is required for constructing the confidence interval.

A:	1.2	1.5	2.3	4.3
B:	4.5	5.7	6.1	8.6

Solution: To verify that the distributions of the piglet weights follow the same distribution other than a shift, side-by-side boxplots as well as comparative dotplots (due to the small sample size) are constructed. The S code that follows can be used to produce graphs similar to those shown in Figure 10.6 on the next page. Based on the graphs in Figure 10.6, it seems reasonable to assume that the underlying distributions are similar in shape.

```
> library(lattice) # Use to get "Trellis" graphs in R
> A <- c(1.2, 1.5, 2.3, 4.3)
> B <- c(4.5, 5.7, 6.1, 8.6)
> n <- length(A)
> m <- length(B)
> r <- c(A, B)
> f <- factor(c(rep("A", n), rep("B", m)))
> graph1 <- bwplot(f~r, xlab="", ylab="Diets", main="Weight Gain")
> graph2 <- dotplot(f~r, xlab="", ylab="Diets", main="Weight Gain")
> print(graph1, split=c(1,1,2,1), more=TRUE)
```

```
> print(graph2, split=c(2,1,2,1), more=FALSE)
```

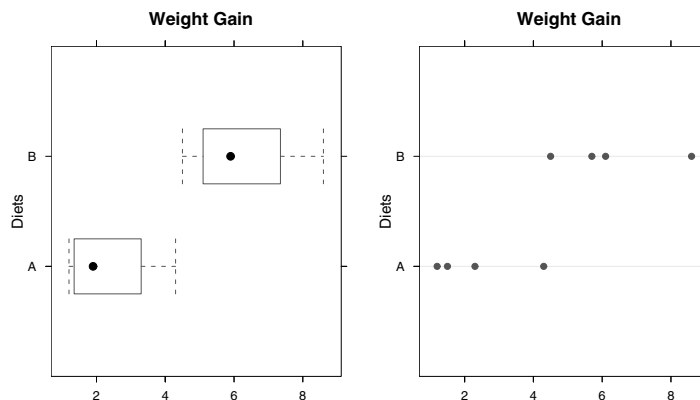


FIGURE 10.6: Side-by-side boxplots as well as comparative dotplots for pig weights for diets *A* and *B*

To find the largest k such that $\mathbb{P}(U < k) \leq \alpha/2$, use the command `pwilcox()`. Since R and S-PLUS use different definitions for `pwilcox`, pay close attention to the code that follows and recall the relationship $U = W - n(n+1)/2$.

```
> pwilcox(1:(n*m), n, m) # R only
> pwilcox( (1+n*(n+1)/2):(n*m+n*(n+1)/2), n, m) # S-PLUS only
 [1] 0.02857143 0.05714286 0.10000000 0.17142857 0.24285714
 [6] 0.34285714 0.44285714 0.55714286 0.65714286 0.75714286
[11] 0.82857143 0.90000000 0.94285714 0.97142857 0.98571429
[16] 1.00000000
```

By visual inspection, one realizes the largest value k such that $\mathbb{P}(U < k) \leq 0.05$ is $k = 2$. That is, the pair $(2, 0.05714286)$ implies a confidence level of $1 - (2)(0.02857143) = 0.9428571$. As an alternative to visual inspection, the appropriate value of k can be found as

```
> pwil <- pwilcox(1:(n*m), n, m) # For R
> which(pwil >= 0.05)[1]
[1] 2
```

Next, the nm differences are generated using the S command `outer()` and the k^{th} and $(nm-k+1)^{\text{st}}$ order statistics from the nm differences are identified. Consequently, a 94.28% confidence interval for the difference in medians is $CI_{0.9428}(\psi_A - \psi_B) = [-7.1, -1.4]$.

```
> k <- 2
> diffs <- matrix(sort(outer(A, B, "-")), byrow=FALSE, nrow=4)
> diffs
      [,1] [,2] [,3] [,4]
[1,] -7.4 -4.6 -3.8 -2.2
[2,] -7.1 -4.5 -3.4 -1.8
[3,] -6.3 -4.3 -3.3 -1.4
[4,] -4.9 -4.2 -3.0 -0.2
```

```
> CL <- 1-2*pwilcox((k-1), n, m)
> CL
[1] 0.9428571
> CI <- c(diffs[k], diffs[n*m-k+1])
> CI
[1] -7.1 -1.4
```

This can be done directly with the argument `conf.int=TRUE` in the `wilcox.test()` function if one is using R:

```
> wilcox.test(A, B, conf.int=TRUE, conf.level=.90) # conf.int=TRUE: only R

      Wilcoxon rank sum test

data:  A and B
W = 0, p-value = 0.02857
alternative hypothesis: true location shift is not equal to 0
90 percent confidence interval:
 -7.1 -1.4
sample estimates:
difference in location
                -4
```

Recall that the achieved confidence level is actually 94.28%. The achieved confidence level is reflected in the output for the function `wilcoxE.test()`. Also note that the statistic `w` in `wilcoxE.test()` is the observed Wilcoxon rank-sum statistic not the Mann-Whitney U statistic reported by R's `wilcox.test()`. The S-PLUS function `wilcox.test()` displays a U statistic that is the Wilcoxon rank-sum statistic.

```
> wilcoxE.test(A, B)

      Wilcoxon Rank Sum Test

data:  A and B
w = 10, p-value = 0.02857
alternative hypothesis: true median is not equal to 0
94.28571 percent confidence interval:
 -7.1 -1.4
sample estimates:
difference in location
                -4
```

10.4.2 Normal Approximation to the Wilcoxon Rank-Sum and Mann-Whitney U -Tests

For moderately sized samples ($n \geq 10$ and $m \geq 10$), the sampling distribution of W can be reasonably approximated with the normal distribution that has mean and standard deviation $n(N+1)/2$ and $\sqrt{nm(N+1)/12}$, respectively. That is, $W \sim N(n(N+1)/2, \sqrt{nm(N+1)/12})$. The standardized test statistic under the assumption that $H_0 : \psi_X - \psi_Y = \delta_0$ is true is

$$Z = \frac{W - \frac{n(N+1)}{2} - \delta_0}{\sqrt{\frac{nm(N+1)}{12}}} \rightsquigarrow N(0, 1). \quad (10.7)$$

See Figure 10.7 for a graph of the Wilcoxon rank-sum distribution for $n = m = 10$ superimposed by a normal distribution with $\mu = n(N + 1)/2 = 105$ and $\sigma = \sqrt{nm(N + 1)/12} = 13.22876$.

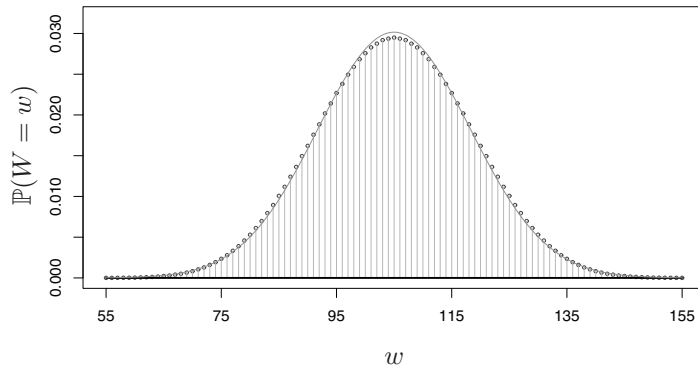


FIGURE 10.7: Graphical representation of the Wilcoxon rank-sum distribution for $n = m = 10$ superimposed by a normal distribution with $\mu = n(N + 1)/2 = 105$ and $\sigma = \sqrt{nm(N + 1)/12} = 13.22876$

The formula for calculating the observed value of the standardized test statistic as well as the three possible alternative hypotheses and their rejection regions are described in Table 10.13 on the next page. If there are tied ranks, the variance of W is reduced to

$$\frac{nm(N + 1)}{12} - \frac{nm}{12N(N - 1)} \sum_{j=1}^g t_j(t_j - 1)(t_j + 1) \quad (10.8)$$

where g denotes the number of tied groups and t_j is the size of tied group j . In (10.8), an untied observation is considered to be a tied group of size one. In the event that no ties exist, $g = N$ and $t_j = 1$ for $j = 1, \dots, N$, which produces a correction factor of zero.

The sampling distribution of U can likewise be reasonably approximated with a normal distribution that has a mean of $nm/2$ and a standard deviation of $\sqrt{nm(N + 1)/12}$. The standardized test statistic under the assumption that $H_0 : \psi_X - \psi_Y = \delta_0$ is true is

$$Z = \frac{U - \frac{nm}{2} - \delta_0}{\sqrt{\frac{nm(N+1)}{12}}} \sim N(0, 1). \quad (10.9)$$

The formula for calculating the observed value of the standardized test statistic as well as the three possible alternative hypotheses and their rejection regions are described in Table 10.14 on the facing page.

A corresponding two-sided confidence interval for the shift in distribution based on (10.9) are the k^{th} and $(nm - k + 1)^{\text{st}}$ ordered differences, where

$$k = 0.5 + \frac{nm}{2} + z_{\alpha/2} \sqrt{\frac{nm(N + 1)}{12}}. \quad (10.10)$$

For a one-sided confidence interval, replace $z_{\alpha/2}$ with z_{α} . Since k is generally not an integer, it can be either rounded or truncated. To obtain a conservative estimate, one should truncate.

Table 10.13: Summary for testing the difference in two medians — normal approximation to the Wilcoxon rank-sum test

Null Hypothesis — $H_0 : \psi_X - \psi_Y = \delta_0$

Standardized Test
Statistic's Value — $z_{obs} = \frac{w \pm 0.5 - n(N+1)/2 - \delta_0}{\sqrt{nm(N+1)/12 - CF}}$

Correction Factor — $CF = \frac{nm}{12N(N-1)} \sum_{j=1}^g t_j(t_j - 1)(t_j + 1)$

Alternative Hypothesis	Rejection Region
$H_1 : \psi_X - \psi_Y < \delta_0$	$z_{obs} < z_\alpha$
$H_1 : \psi_X - \psi_Y > \delta_0$	$z_{obs} > z_{1-\alpha}$
$H_1 : \psi_x - \psi_Y \neq \delta_0$	$ z_{obs} > z_{1-\alpha/2}$
<p>Note: The quantity ± 0.5 in the numerator of z_{obs} is the continuity correction. When $H_1 : \psi_X - \psi_Y < \delta_0$, the quantity $+0.5$ is used. When $H_1 : \psi_X - \psi_Y > \delta_0$, the quantity -0.5 is used. When $H_1 : \psi_x - \psi_Y \neq \delta_0$, use $+0.5$ if $w < n(N + 1)/2$ and -0.5 if $w > n(N + 1)/2$.</p>	

Table 10.14: Summary for testing the difference in two medians — normal approximation to the Mann-Whitney U -Test

Null Hypothesis — $H_0 : \psi_X - \psi_Y = \delta_0$

Standardized Test
Statistic's Value — $z_{obs} = \frac{u \pm 0.5 - nm/2 - \delta_0}{\sqrt{nm(N+1)/12 - CF}}$

Correction Factor — $CF = \frac{nm}{12N(N-1)} \sum_{j=1}^g t_j(t_j - 1)(t_j + 1)$

Alternative Hypothesis	Rejection Region
$H_1 : \psi_X - \psi_Y < \delta_0$	$z_{obs} < z_\alpha$
$H_1 : \psi_X - \psi_Y > \delta_0$	$z_{obs} > z_{1-\alpha}$
$H_1 : \psi_x - \psi_Y \neq \delta_0$	$ z_{obs} > z_{1-\alpha/2}$
<p>Note: The quantity ± 0.5 in the numerator of z_{obs} is the continuity correction. When $H_1 : \psi_X - \psi_Y < \delta_0$, the quantity $+0.5$ is used. When $H_1 : \psi_X - \psi_Y > \delta_0$, the quantity -0.5 is used. When $H_1 : \psi_x - \psi_Y \neq \delta_0$, use $+0.5$ if $u < nm/2$ and -0.5 if $u > nm/2$.</p>	

Example 10.9 ▷ *Wilcoxon Rank-Sum Test: Swim Times* ◁ Thirty-two division I swimmers from the same swim team agree to participate in a year-long study to determine whether high (30%) fat diets produce greater improvements in swim times than the standard low (10%) fat diets. Times for the 32 swimmers for the 200 yard individual medley were taken right after the swimmers' conference meet. The swimmers were randomly assigned to follow one of the diets. The group on diet 1 followed a low-fat diet the entire year but lost two swimmers along the way. The group on diet 2 followed the high fat diet the entire year and also lost two swimmers. Times for the 200 yard individual medley were taken one year later for the remaining 28 swimmers. The swimmers' improvements in seconds for both diets are presented in Table 10.15 on the next page and stored in data frame `Swimtimes`, where x_i represents the time improvement in seconds for swimmers on high fat diet (`highfat`) and y_i represents the time improvement in seconds for swimmers on low-fat diet (`lowfat`).

- (a) Verify that the time improvement distributions are similar in shape.
- (b) Test whether the median difference for improvements in swim times is different from zero using a significance level of $\alpha = 0.10$ with the normal approximation to the Wilcoxon rank-sum test and the normal approximation to the Mann-Whitney U -test.
- (c) Use the function `wilcox_test()` from the `coin` package, which can be downloaded from your nearest CRAN mirror at <http://cran.r-project.org/mirrors.html> to report the exact p -value and the 90% confidence interval for the hypothesis in (b). According to the documentation, this function computes exact conditional (on the data) p -values and quantiles using the shift-algorithm by Streitberg and Röhmel for both tied and untied samples.
- (d) Construct a confidence interval with confidence level of at least 90% using the normal approximation to find k .

Solution: The answers are as follows:

- (a) To use the Wilcoxon rank-sum test, the time improvement distributions must be similar in shape. A comparative boxplot of time improvements for low-fat and high fat diets is found in Figure 10.8 on the facing page. Since the comparative boxplot does appear to show the same underlying distribution for time improvements for swimmers eating both diets, it is legitimate to proceed with a Wilcoxon rank-sum test or the Mann-Whitney U -test.
- (b) Use the five-step procedure to test if the median difference for improvements in swim times for high and low-fat diets is different from zero.

Step 1: **Hypotheses** — The null and alternative hypotheses to test if the median difference for improvements in swim times for high and low-fat diets is different from zero are

$$H_0 : \psi_X - \psi_Y = 0 \text{ versus } H_1 : \psi_X - \psi_Y \neq 0.$$

Step 2: **Test Statistic** —

Wilcoxon Rank-Sum Test

The test statistic chosen is W , where the observed value is

$$w = 248.$$

Mann-Whitney Test

The test statistic chosen is U , where the observed value is

$$\begin{aligned} u &= w - n(n + 1)/2 \\ &= 248 - 14(15)/2 = 143. \end{aligned}$$

Table 10.15: Sorted improvements in swim times in seconds for high (x) and low (y) fat diets, where rank refers to the rank of the data point in the combined sample of x and y data points (**Swimtimes**)

	y_i	rank(y_i)	x_i	rank(x_i)
Tied Rank	0.18	8.5	0.18	8.5
	-0.79	2.0	0.38	10.0
	-0.49	3.0	0.56	11.0
	-0.37	4.0	0.65	12.0
	-0.20	5.0	0.84	13.0
	-0.15	6.0	1.58	20.0
	0.02	7.0	0.89	16.0
	-0.87	1.0	1.18	18.0
Tied Rank	0.87	14.5	0.87	14.5
	0.98	17.0	2.03	22.0
	1.42	19.0	3.53	27.0
	1.71	21.0	4.33	28.0
	3.52	26.0		
Tied Ranks	2.66	24.0	2.66	24.0
			2.66	24.0

$w = 248$

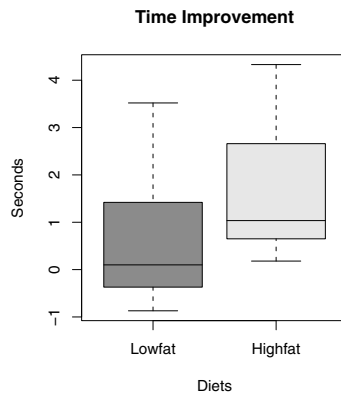


FIGURE 10.8: Comparative boxplot for improvements in swim times for high and low-fat diets

Step 3: **Rejection Region Calculations** — Because both standardized test statistics are distributed approximately $N(0, 1)$ and H_1 is a two-sided hypothesis, the rejection region is $|z_{obs}| > z_{1-0.10/2} = 1.645$.

Because there are three groups of ties ($g = 3$), where the sizes of the tied groups are 2, 2, and 3, the correction factor is

$$\begin{aligned} CF &= \frac{nm}{12N(N-1)} \sum_{j=1}^3 t_j(t_j-1)(t_j+1) \\ &= \frac{(14)(14)}{12(28)(28-1)} \times [2(2-1)(2+1) \\ &\quad + 2(2-1)(2+1) + 3(3-1)(3+1)] \\ &= \frac{7}{324} [6 + 6 + 24] \\ &= 7/9 \end{aligned}$$

Wilcoxon Rank-Sum Test

The value of the standardized test statistic is

$$\begin{aligned} z_{obs} &= \frac{w \pm 0.5 - n(N+1)/2 - \delta_0}{\sqrt{nm(N+1)/12 - CF}} \\ &= \frac{248 - 0.5 - 14(28+1)/2 - 0}{\sqrt{(14)(14)(28+1)/12 - 7/9}} \\ &= 2.046353 \end{aligned}$$

Mann-Whitney Test

The value of the standardized test statistic is

$$\begin{aligned} z_{obs} &= \frac{u \pm 0.5 - nm/2 - \delta_0}{\sqrt{nm(N+1)/12 - CF}} \\ &= \frac{143 - 0.5 - (14)(14)/2 - 0}{\sqrt{(14)(14)(28+1)/12 - 7/9}} \\ &= 2.046353 \end{aligned}$$

Step 4: **Statistical Conclusion** — The φ -value is $2\mathbb{P}(Z \geq 2.046353) = 0.04072$.

- I. From the rejection region, reject H_0 because $z_{obs} = 2.046353$ is more than 1.645.
- II. From the φ -value, reject H_0 because the φ -value = 0.04072 is less than 0.10.

Reject H_0 .

Step 5: **English Conclusion** — There is sufficient evidence to suggest that the median time improvements are different for swimmers eating high fat and low-fat diets.

The φ -value can be computed with the S command `wilcox.test(highfat, lowfat)`. In the presence of ties, the S function `wilcox.test()` automatically uses the normal approximation to the distribution of U (R) and W (S-PLUS), as well as applying a correction factor for the variances of U and W and an appropriate continuity correction factor to agree with the formula for the standardized test statistic given in Tables 10.13 and 10.14 on page 431. Output for both R and S-PLUS is provided. R does not report the value of the standardized test statistic but does use the value of the standardized test statistic to compute the φ -value. S-PLUS does report the value of the standardized test statistic (see the output that follows). The φ -value from the output is 0.04072, the exact value found for a z_{obs} value of 2.046353 in step 4.

Output for R:

```
> attach(Swimtimes)
> wilcox.test(highfat, lowfat)

      Wilcoxon rank sum test with continuity correction

data:  highfat and lowfat
W = 143, p-value = 0.04072
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(highfat, lowfat) :
  cannot compute exact p-value with ties
```

Output for S-PLUS:

```
> wilcox.test(highfat, lowfat)
Warning messages:
  cannot compute exact p-value with ties in:
  wil.rank.sum(x, y, alternative, exact, correct)
```

Wilcoxon rank-sum test

```
data:  highfat and lowfat
rank-sum normal statistic with correction Z = 2.0464, p-value = 0.0407
alternative hypothesis: mu is not equal to 0
```

(c) From the output of `wilcox_test()`, the p -value is 0.03818 and the 90% confidence interval for the difference in medians is $CI_{0.90}(\psi_X - \psi_Y) = [0.31, 1.68]$.

```
> library(coin) # needed for wilcox_test()
> GR <- factor(c(rep("lowfat",14), rep("highfat",14)))
> wilcox_test(c(lowfat, highfat)~GR, distribution="exact",
+ conf.int=TRUE, conf.level=.90) # Only R
```

Exact Wilcoxon Mann-Whitney Rank Sum Test

```
data:  c(lowfat, highfat) by GR (highfat, lowfat)
Z = 2.0693, p-value = 0.03818
alternative hypothesis: true mu is not equal to 0
90 percent confidence interval:
 0.31 1.68
sample estimates:
difference in location
      1.02
```

Note that the reported standardized test statistic computed with `wilcox_test()` does not use a continuity correction.

(d) The $x_i - y_j$ differences are stored in `diffs`. Using (10.10) on page 430, k is calculated to be 62. The k^{th} difference is determined to be 0.31 and the $nm - k + 1^{\text{st}}$ difference is 1.68. Therefore, the 90% confidence interval, $CI_{0.90}(\psi_X - \psi_Y)$, is $[0.31, 1.68]$.

```

> n <- length(highfat)
> m <- length(lowfat)
> N <- n + m
> diffs <- sort(outer(highfat, lowfat, "-"))
> k <- 0.5 + n*m/2 + qnorm(.05)*sqrt(n*m*(N+1)/12) # k for 90% CI
> k <- floor(k)
> k
[1] 62
> CI <- c(diffs[k], diffs[n*m-k+1]) #90% CI
> CI
[1] 0.31 1.68

```

■

10.5 The Kruskal-Wallis Test

The Kruskal-Wallis test is an extension of the Wilcoxon rank-sum/Mann-Whitney U -test for two independent samples (covered in Section 10.4) to the situation with a mutually independent samples. As with most statistical procedures, independence is preserved by using random samples. The design structure of this problem is often called a completely randomized design. The null hypothesis is that the a populations are identical. Like the Wilcoxon rank-sum/Mann-Whitney U -test, the only assumption the Kruskal-Wallis test requires is that the a populations be continuous. The null and alternative hypotheses are written

$$\begin{aligned}
 H_0 : F_1(x) = F_2(x) = \cdots = F_a(x) \text{ for all } x & \quad \text{versus} \\
 H_1 : F_i(x) \neq F_j(x) \text{ for at least one pair } (i, j) \text{ and some } x.
 \end{aligned}
 \tag{10.11}$$

Because the underlying distributions of the a populations are assumed to be identical in the null hypothesis, this test can apply to means, medians, or any other quantile and the null and alternative hypotheses are often expressed in terms of the population medians as

$$H_0 : \psi_1 = \psi_2 = \cdots = \psi_a \text{ versus } H_1 : \psi_i \neq \psi_j \text{ for at least one pair } (i, j)
 \tag{10.12}$$

To test the null hypothesis, all n_1, n_2, \dots, n_a observations are pooled into a single column and ranked from 1 to $N = \sum_{i=1}^a n_i$. The standardized test statistic that both R and S-PLUS use with the Kruskal-Wallis test via the function `kruskal.test()` is

$$H = \frac{12 \sum_{i=1}^a n_i (\bar{R}_i - \bar{R}_\bullet)^2}{N(N+1)}
 \tag{10.13}$$

where n_i is the number of observations in the i^{th} treatment/group, \bar{R}_i is the average of the ranks in the i^{th} treatment/group, and \bar{R}_\bullet is the average of all of the ranks. When ties are present, an adjusted standardized test statistic denoted as H' is also calculated and reported. The adjusted statistic H' is defined as

$$H' = \frac{H}{f_c} = \frac{H}{1 - \frac{\sum_{j=1}^r (t_j^3 - t_j)}{N^3 - N}}
 \tag{10.14}$$

where t_j is the number of times a given rank was tied in the combined sample of size N and r is the number of ranks in the combined sample of size N that were tied. Provided each

$n_i \geq 5$, the sampling distributions of H_{obs} and H' are both approximately chi-square random variables with $a - 1$ degrees of freedom (χ_{a-1}^2). More specifically, when H_0 is true, the statistic H has, as $\min(n_1, \dots, n_a)$ tends to infinity, an asymptotic χ_{a-1}^2 distribution. The arguments for `kruskal.test()` differ for R and S-PLUS, yet the test statistic is computed according to (10.13) and (10.14) when ties are present in the data.

Example 10.10 ▷ *Kruskal-Wallis Test: Free Throws* ◁ An elementary school gym teacher is interested in evaluating the effectiveness of four free throw teaching techniques. The gym teacher randomly assigns the 80 students to one of four groups with 20 students per group. After two months, every member of the groups shoots 10 free throws, and the gym teacher records the results. The number of successful free throws each student shoots in each of the four groups is presented in Table 10.16. Use the free throw results to decide if differences exist among teaching methods at the $\alpha = 0.05$ level.

Table 10.16: Number of successful free throws

Method	Data																			
Method1	6	1	2	0	0	1	1	3	1	2	1	2	4	2	1	1	1	3	7	1
Method2	3	2	1	2	1	6	2	1	1	2	1	1	2	3	2	2	3	2	5	2
Method3	2	1	2	3	2	2	4	3	2	3	2	5	1	1	3	7	6	2	2	2
Method4	2	1	1	3	1	2	1	6	1	1	0	1	1	1	1	2	2	1	5	4

Solution: The five-step procedure is used and explained to determine if differences exist among teaching methods. Before proceeding, first examine side-by-side boxplots for free throws made grouped by teaching method. Based on the boxplots and the density plots in Figure 10.9, it seems reasonable to assume that all a populations are similar in shape.

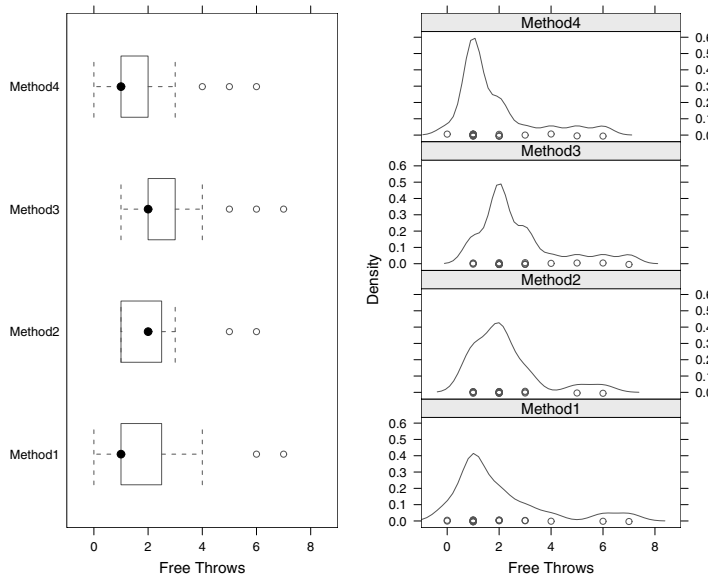


FIGURE 10.9: Boxplots and density plots of free throw teaching results


```

> library(lattice)      # Only for R
> Method1 <- c(6,1,2,0,0,1,1,3,1,2,1,2,4,2,1,1,1,3,7,1)
> Method2 <- c(3,2,1,2,1,6,2,1,1,2,1,1,2,3,2,2,3,2,5,2)
> Method3 <- c(2,1,2,3,2,2,4,3,2,3,2,5,1,1,3,7,6,2,2,2)
> Method4 <- c(2,1,1,3,1,2,1,6,1,1,0,1,1,1,1,2,2,1,5,4)
> n1 <- length(Method1); n2 <- length(Method2)
> n3 <- length(Method3); n4 <- length(Method4)
> NumberFT <- c(Method1, Method2, Method3, Method4)
> g <- as.factor(c(rep("Method1", n1), rep("Method2", n2),
+ rep("Method3", n3), rep("Method4", n4)))      # Methods
> # Equivalently, g could be created using
> g <- factor(rep(c("Method1", "Method2", "Method3", "Method4"), rep(20,4)))
> A <- bwplot(g~NumberFT, xlab="Free Throws", xlim=c(-1,9))
> B <- densityplot(~NumberFT|g, layout=c(1,4),
+ xlab="Free Throws", xlim=c(-1,9))
> print(A, split=c(1,1,2,1), more=TRUE)
> print(B, split=c(2,1,2,1), more=FALSE)

```

Step 1: **Hypotheses** — The hypotheses to test equality of the F_i for $i = 1, \dots, a$ distributions are $H_0 : F_1(x) = F_2(x) = F_3(x) = F_4(x)$ for all x versus $H_1 : F_i(x) \neq F_j(x)$ for at least one pair (i, j) and some x

Step 2: **Test Statistic** — The test statistic R_i is used to evaluate the null hypothesis. Under the assumption that H_0 is true, the standardized test statistic

$$H = \frac{12 \sum_{i=1}^a n_i (\bar{R}_i - \bar{R}_\bullet)^2}{N(N+1)} \rightsquigarrow \chi_{a-1}^2$$

Step 3: **Rejection Region Calculations** — The rejection region is $H_{\text{obs}} > \chi_{95;3}^2 = 7.815$. The number of free throws completed in each of the methods is combined into a single population and ranked among the 80 observations. Table 10.17 on the next page shows the actual free throws with their ranks among the 80 observations. The value of H_{obs} is calculated as

$$\begin{aligned}
 H_{\text{obs}} &= \frac{12 \sum_{i=1}^a n_i (\bar{R}_i - \bar{R}_\bullet)^2}{N(N+1)} \\
 &= \frac{12}{(80 \times 81)} \times \left\{ (20 \times (35.05 - 40.50))^2 + (20 \times (42.875 - 40.50))^2 + \right. \\
 &\quad \left. (20 \times (50.825 - 40.50))^2 + (20 \times (33.250 - 40.50))^2 \right\} \\
 &= 7.20412
 \end{aligned}$$

Table 10.17: Actual free throws with ranks among all free throws

Meth1	RankM1	Meth2	RankM2	Meth3	RankM3	Meth4	RankM4	Total
6	76.5	3	63.5	2	45.5	2	45.5	
1	18.0	2	45.5	1	18.0	1	18.0	
2	45.5	1	18.0	2	45.5	1	18.0	
0	2.0	2	45.5	3	63.5	3	63.5	
0	2.0	1	18.0	2	45.5	1	18.0	
1	18.0	6	76.5	2	45.5	2	45.5	
1	18.0	2	45.5	4	70.0	1	18.0	
3	63.5	1	18.0	3	63.5	6	76.5	
1	18.0	1	18.0	2	45.5	1	18.0	
2	45.5	2	45.5	3	63.5	1	18.0	
1	18.0	1	18.0	2	45.5	0	2.0	
2	45.5	1	18.0	5	73.0	1	18.0	
4	70.0	2	45.5	1	18.0	1	18.0	
2	45.5	3	63.5	1	18.0	1	18.0	
1	18.0	2	45.5	3	63.5	1	18.0	
1	18.0	2	45.5	7	79.5	2	45.5	
1	18.0	3	63.5	6	76.5	2	45.5	
3	63.5	2	45.5	2	45.5	1	18.0	
7	79.5	5	73.0	2	45.5	5	73.0	
1	18.0	2	45.5	2	45.5	4	70.0	
Means:	35.050		42.875		50.825		33.250	40.500

The adjusted test statistic H'_{obs} is calculated as

$$\begin{aligned}
 H'_{\text{obs}} &= \frac{H_{\text{obs}}}{f_c} = \frac{H_{\text{obs}}}{1 - \frac{\sum_{j=1}^r (t_j^3 - t_j)}{N^3 - N}} \\
 &= \frac{7.20412}{1 - \left\{ \frac{(3^3-3)+(29^3-29)+(26^3-26)+(10^3-10)+(3^3-3)+(3^3-3)+(4^3-4)+(2^3-2)}{80^3-80} \right\}} \\
 &= \frac{7.204120}{0.9159283} = 7.865376
 \end{aligned}$$

Step 4: **Statistical Conclusion** — The φ -value for the standardized test statistic without adjustment for ties (H) and the standardized test statistic adjusted for ties (H') are calculated as $\mathbb{P}(\chi_3^2 \geq 7.204) = 0.0656$ and $\mathbb{P}(\chi_3^2 \geq 7.86) = 0.0488$, respectively. φ -values such as 0.065 and 0.0488 indicate that observing values as extreme or more than 7.20 or 7.86 when the null hypothesis is true are fairly unlikely.

- I. From the rejection region, reject H_0 since $H'_{\text{obs}} = 7.86 > \chi_{.95;3}^2 = 7.81$.
- II. From the φ -value, reject H_0 because the φ -value = 0.0488 is less than 0.05.

Reject H_0 .

Step 5: **English Conclusion** — There is statistical evidence to suggest differences exist among the distributions for the four free throw teaching methods.

To compute the rejection region, the value of the standardized test statistic (Hobs), and the value of the standardized test statistic corrected for ties (Hc) with S, enter

```
> RR <- qchisq(.95,3) # Rejection region
> RR
[1] 7.814728
> RKs <- rank(NumberFT) # Ranks for all
> MRKs <- unstack(RKs, RKs~g) # Only R not S-PLUS
> MRKs[1:5,] # Show first five rows
  Method1 Method2 Method3 Method4
1   76.5   63.5   45.5   45.5
2   18.0   45.5   18.0   18.0
3   45.5   18.0   45.5   18.0
4    2.0   45.5   63.5   63.5
5    2.0   18.0   45.5   18.0
> RK <- apply(MRKs,2, mean) # Treatment ranks
> names(RK) <- c("MRKT1","MRKT2","MRKT3","MRKT4")
> RK
  MRKT1 MRKT2 MRKT3 MRKT4
35.050 42.875 50.825 33.250
> MRK <- mean(RK) # Overall mean rank
> MRK
[1] 40.5
> N <- length(RKs)
> Hobs <- 12*(n1*(RK[1] - MRK)^2 + n2*(RK[2] - MRK)^2
+ + n3*(RK[3] - MRK)^2 + n4*(RK[4] - MRK)^2)/(N*(N+1))
> names(Hobs) <- "statistic"
> Hobs
statistic
 7.20412
> tj <- table(RKs)
> tj
RKs
  2  18 45.5 63.5  70  73 76.5 79.5
  3  29  26  10  3  3  4  2
> CF <- 1-(sum(tj^3 - tj)/(N^3-N)) # correction factor
> Hc <- Hobs/CF # corrected statistic
> hs <- c(Hobs, Hc)
> hs
statistic statistic
 7.204120 7.865376
> pval <- 1-pchisq(hs,3)
> names(pval) <- c("p.value","p.value")
> pval
  p.value  p.value
0.06566864 0.04887747
```

To find the standardized test statistic corrected for ties and its corresponding ϕ -value with the function `kruskal.test()`, enter

```
> kruskal.test(NumberFT~g) # For S-PLUS change ~ to ,
```

Kruskal-Wallis rank sum test

```
data: NumberFT by g
Kruskal-Wallis chi-squared = 7.8654, df = 3,
p-value = 0.04888
```

■

Distribution of H The exact distribution of H can be obtained using the fact that under H_0 , all possible $(\sum_{i=1}^a n_i)! / (\prod_{i=1}^a n_i!)$ assignments of n_1 ranks to the treatment 1 observations, n_2 ranks to the treatment 2 observations, \dots , n_a ranks to the treatment a observations are equally likely; however, there are practical and computational limits on the range of tables that can be constructed. Consider that Example 10.10 on page 437 has $80! / (20! \cdot 20! \cdot 20! \cdot 20!) = 4.895203 \times 10^{46}$ possible rank assignments. Consequently, the distribution of H is generally approximated with a χ_{a-1}^2 distribution. Under the null hypothesis, the n_i ranks in sample i are randomly selected from the set $\{1, 2, \dots, \sum_{i=1}^a n_i = N\}$. That is, the ranks in sample i are drawn without replacement from the finite populations of N ranks.

For a finite population, it can be shown that

$$E[\bar{R}_i] = \frac{N+1}{2} \quad \text{and} \quad \text{Var}[\bar{R}_i] = \frac{(N+1)(N-n_i)}{12n_i}.$$

Provided the $\min(n_i)$ is sufficiently large,

$$Z_i = \frac{\bar{R}_i - \frac{N+1}{2}}{\sqrt{\frac{(N+1)(N-n_i)}{12n_i}}} \sim N(0, 1) \quad (10.15)$$

by the Central Limit Theorem. It then follows that $Z_i^2 \sim \chi_1^2$. Although the Z_i s are not independent, when H_0 is true, the statistic

$$H = \sum_{i=1}^a \frac{N-n_i}{N} Z_i^2 = \sum_{i=1}^a \frac{12n_i [\bar{R}_i - \frac{N+1}{2}]^2}{N(N+1)} \quad (10.16)$$

has, as $\min\{n_1, n_2, \dots, n_a\}$ tends to infinity, an asymptotic χ_{a-1}^2 distribution. When the null hypothesis is rejected, one can compare any two groups by calculating

$$Z_{ij\text{obs}} = \frac{|\bar{R}_i - \bar{R}_j|}{\sqrt{\left(\frac{N(N+1)}{12}\right) \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \quad (10.17)$$

and declaring treatments i and j significantly different when $Z_{ij\text{obs}} > Z_{1-\alpha/[a(a-1)]}$. By dividing $\alpha/2$ by $a(a-1)/2$, the number of pairwise comparisons, the overall significance level is appropriately adjusted.

Example 10.10 on page 437 rejected the null hypothesis of equal distributions and concluded that at least two of the four methods have different distributions. The next step is to decide which one of the four methods the gym instructor should use in teaching students to shoot free throws. Using (10.17) with an $\alpha = 0.20$, methods 1 and 4 are declared to be significantly different from method 3 since $Z_{13\text{obs}} = 2.15 > Z_{1-\alpha/(a(a-1))} = 2.13$ and $Z_{34\text{obs}} = 2.39 > Z_{1-\alpha/(a(a-1))} = 2.13$. In this case, the probability that all the statements

are correct is $1 - \alpha = 0.8$. The gym instructor should stop using methods 1 and 4. If the instructor had to pick only one method to use, and all other factors were the same, he/she should use method 3 since it is statistically better than method 1 and method 4. Although there is no statistical difference between methods 2 and 3, method 2 is not statistically better than method 1 or method 4. Code to compute the multiple comparisons according to (10.17) is

```
> a <- 4                                     # Four methods
> Method1 <- c(6,1,2,0,0,1,1,3,1,2,1,2,4,2,1,1,1,3,7,1)
> Method2 <- c(3,2,1,2,1,6,2,1,1,2,1,1,2,3,2,2,3,2,5,2)
> Method3 <- c(2,1,2,3,2,2,4,3,2,3,2,5,1,1,3,7,6,2,2,2)
> Method4 <- c(2,1,1,3,1,2,1,6,1,1,0,1,1,1,1,2,2,1,5,4)
> n1 <- length(Method1); n2 <- length(Method2)
> n3 <- length(Method3); n4 <- length(Method4)
> NumberFT <- c(Method1, Method2, Method3, Method4)
> N <- length(NumberFT)
> RKs <- rank(NumberFT)                     # Ranks for all
> MRKs <- unstack(RKs, RKs~g)              # Only R not S-PLUS
> RK <- apply(MRKs, 2, mean)                # Treatment ranks
> names(RK) <- c("MRKT1", "MRKT2", "MRKT3", "MRKT4")
> alpha <- 0.20
> Z12 <- abs(RK[1]-RK[2])/sqrt((N*(N+1)/12)*(1/n1 + 1/n2))
> Z13 <- abs(RK[1]-RK[3])/sqrt((N*(N+1)/12)*(1/n1 + 1/n3))
> Z14 <- abs(RK[1]-RK[4])/sqrt((N*(N+1)/12)*(1/n1 + 1/n4))
> Z23 <- abs(RK[2]-RK[3])/sqrt((N*(N+1)/12)*(1/n2 + 1/n3))
> Z24 <- abs(RK[2]-RK[4])/sqrt((N*(N+1)/12)*(1/n2 + 1/n4))
> Z34 <- abs(RK[3]-RK[4])/sqrt((N*(N+1)/12)*(1/n3 + 1/n4))
> Zij <- round(c(Z12, Z13, Z14, Z23, Z24, Z34), 2)
> names(Zij) <- c("Z12", "Z13", "Z14", "Z23", "Z24", "Z34")
> CV <- round(qnorm(1- alpha/(a*(a-1))), 2)
> Zij
  Z12  Z13  Z14  Z23  Z24  Z34
1.06 2.15 0.24 1.08 1.31 2.39
> CV
[1] 2.13
> which(Zij > CV)
Z13 Z34
  2   6
```

10.6 Friedman Test for Randomized Block Designs

In Section 10.5, the Kruskal-Wallis rank test for several independent samples was introduced as an extension of the Wilcoxon rank-sum/Mann-Whitney U -test for two independent samples introduced in Section 10.4. In this section, the problem of analyzing related samples is examined. The design structure of the problems addressed in this section is often referred to as a randomized complete block design. In this type of design, there are b blocks and $k \geq 2$ treatments, and the test is designed to detect differences among the k treatments. In this type of scenario, observations are arranged in blocks, which are groups of k experimental units similar to each other in some important characteristic. The rationale behind using

a block is to reduce the error of the experiment as much as possible by grouping similar units so that the remaining differences will be largely due to the treatments. The use of “blocks” comes from some of the earliest experimental designs in agriculture where fields were divided in “blocks.”

In a randomized complete block design (RCBD), experimental units are assigned to blocks, and then treatments are randomly assigned to the units within the blocks. To analyze a RCBD with Friedman’s test, ranks are assigned to the observations within each block. The ranked observations are denoted R_{ij} , $i = 1, \dots, b$, $j = 1, \dots, k$. A representation of the ranked data from a RCBD is shown in Table 10.18.

Table 10.18: A representation of the ranked data from a randomized complete block design

	1	2	\dots	k	Row Totals
1	R_{11}	R_{12}	\dots	R_{1k}	$k(k + 1)/2$
2	R_{21}	R_{22}	\dots	R_{2k}	$k(k + 1)/2$
⋮	⋮			⋮	⋮
b	R_{b1}	R_{b2}	\dots	R_{bk}	$k(k + 1)/2$
Column Totals:	R_1	R_2	\dots	R_k	$bk(k + 1)/2$

The assumptions required to apply Friedman’s test are the same as those required for the Kruskal-Wallis test; namely, all populations sampled are continuous and identical, except possibly for location. The null hypothesis is that the populations all have the same location. Typically, the null hypothesis of no difference among the k treatments is written in terms of the medians as $H_0 : \psi_1 = \psi_2 = \dots = \psi_k$. Although the distribution under the null hypothesis could be enumerated, it is not practical to do so as there are a total of $(k!)^b$ distinguishable sets of entries in a $b \times k$ table. The Friedman statistic S is

$$S = \left[\frac{12}{bk(k + 1)} \sum_{j=1}^k R_j^2 \right] - 3b(k + 1), \tag{10.18}$$

where R_j is the sum of ranks for each treatment, where ranks were assigned within each block. The statistic S has an asymptotic χ_{k-1}^2 distribution as b tends to infinity. For $b > 7$, numerical comparisons have shown χ_{k-1}^2 to be a reasonable approximation to the distribution of S (Gibbons and Chakraborti, 2003). When ties are present in the ranks, S is replaced with the quantity S' :

$$S' = \frac{12 \sum_{j=1}^k R_j^2 - 3b^2k(k + 1)^2}{bk(k + 1) - \frac{1}{k-1} \sum_{i=1}^b \left\{ \left(\sum_{j=1}^{g_i} t_{ij}^3 \right) - k \right\}} \tag{10.19}$$

where g_i denotes the number of tied groups in the i^{th} block and t_{ij} is the size of the j^{th} tied group in the i^{th} block. Note that when there are no ties in the blocks, the quantity $\frac{1}{k-1} \sum_{i=1}^b \left\{ \left(\sum_{j=1}^{g_i} t_{ij}^3 \right) - k \right\} = 0$ and S' reduces to S . The null hypothesis is rejected at the α level of significance whenever $S'_{\text{obs}} > \chi_{1-\alpha; k-1}^2$. When the null hypothesis is rejected, one can declare treatments i and j significantly different when $ZR_{ij \text{ obs}} > Z_{\alpha/[k(k-1)]}$, where

$ZR_{ij\text{obs}}$ is defined as

$$ZR_{ij\text{obs}} = \frac{|R_i - R_j|}{\sqrt{\left(\frac{bk(k+1)}{6}\right)}}. \quad (10.20)$$

Typical values of α when performing multiple comparisons are often as large as 0.20 due to the large number of comparisons.

Example 10.11 ▷ *Friedman Test: Body Fat* ◁ The body fat of 78 high school wrestlers was measured using three separate techniques and the results are stored in the data frame `HSwrestler`. The techniques used were hydrostatic weighing (HWFAT), skin fold measurements (SKFAT), and the Tanita body fat scale (TANFAT). Do the three methods of recording body fat have equal medians? Use a significance level of $\alpha = 0.05$ to reach your conclusion. If the null hypothesis of equal medians is rejected, determine which treatments are significantly different using an overall experiment-wise error rate of $\alpha = 0.20$.

Solution: Each wrestler in this scenario acts as a block. This particular design structure is also known as a repeated measures design. Before testing the null hypothesis of equal medians, a few graphs are created to verify the assumption of equal shaped populations.

```
> attach(HSwrestler)
> HSwrestler[1:5,]
  AGE   HT   WT ABS TRICEPS SUBSCAP HWFAT TANFAT SKFAT
1  18 65.75 133.6   8     6    10.5 10.71   11.9  9.80
2  15 65.50 129.0  10     8     9.0  8.53   10.0 10.56
3  17 64.00 120.8   6     6     8.0  6.78    8.3  8.43
4  17 72.00 145.0  11    10    10.0  9.32    8.2 11.77
5  17 69.50 299.2  54    42    37.0 41.89   41.6 41.09
> FAT <- c(HWFAT, TANFAT, SKFAT)
> GROUP <- factor(rep(c("HWFAT", "TANFAT", "SKFAT"), rep(78, 3)))
> BLOCK <- factor(rep(1:78, 3)) # used later
> library(lattice)
> A <- bwplot(GROUP~FAT, xlab="% Fat")
> B <- densityplot(~FAT|GROUP, layout=c(1,3), xlab="% Fat")
> print(A, split=c(1,1,2,1), more=TRUE)
> print(B, split=c(2,1,2,1), more=FALSE)
```

Based on the boxplots and the density plots in Figure 10.10 on the next page, it seems reasonable to assume that the distributions of body fat for the three treatment groups are similar in shape.

Step 1: **Hypotheses** — The hypotheses to test no difference among the k treatments are $H_0 : \psi_1 = \psi_2 = \dots = \psi_k$ versus $H_1 : \psi_i \neq \psi_j$ for at least one pair (i, j) .

Step 2: **Test Statistic** — The test statistic S is used to evaluate the null hypothesis. Under the assumption that H_0 is true, the standardized test statistic

$$S = \left[\frac{12}{bk(k+1)} \sum_{j=1}^k R_j^2 \right] - 3b(k+1) \sim \chi_{b-1}^2$$

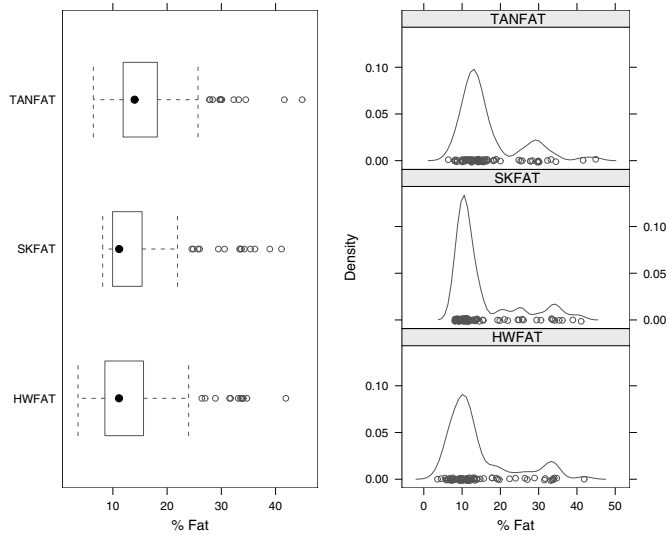


FIGURE 10.10: Comparative boxplots and density plots for hydrostatic weighing (HWFAT), skin fold measurements (SKFAT), and the Tanita body fat scale (TANFAT)

Step 3: **Rejection Region Calculations** — The rejection region is $S_{\text{obs}} > \chi^2_{.95;2} = 5.99$. The first six wrestlers' body fat as measured by the three techniques and their corresponding ranks are shown in Table 10.19.

Table 10.19: The first six wrestlers' body fat as measured by the three techniques and their corresponding ranks

Wrestler	Measurement			Rank		
	HWFAT	TANFAT	SKFAT	HWFAT	TANFAT	SKFAT
1	10.71	11.9	9.80	2	3	1
2	8.53	10.0	10.56	1	2	3
3	6.78	8.3	8.43	1	2	3
4	9.32	8.2	11.77	2	1	3
5	41.89	41.6	41.09	3	2	1
6	34.03	29.9	29.45	3	2	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
				$R_1 = 128$	$R_2 = 187$	$R_3 = 153$

The value of S_{obs} is calculated as

$$\begin{aligned}
 S_{\text{obs}} &= \left[\frac{12}{bk(k+1)} \sum_{j=1}^k R_j^2 \right] - 3b(k+1) \\
 &= \left[\frac{12}{78 \cdot 3(3+1)} (128^2 + 187^2 + 153^2) \right] - 3 \cdot 78(3+1) \\
 &= 22.48718
 \end{aligned}$$

Step 4: **Statistical Conclusion** — The φ -value for the standardized test statistic (S_{obs}) is calculated as $\mathbb{P}(\chi_2^2 \geq 22.48718) = 1.309095 \times 10^{-5}$, indicating that observing values as extreme or more than 22.48718 when the null hypothesis is true is very unlikely.

- I. From the rejection region, reject H_0 since $S_{\text{obs}} = 22.48718 > \chi_{.95;2}^2 = 5.99$.
- II. From the φ -value, reject H_0 because the φ -value = 1.309095×10^{-5} is less than 0.05.

Reject H_0 .

Step 5: **English Conclusion** — There is statistical evidence to suggest differences exist among the three methods used to measure body fat.

To compute the rejection region, the value of the standardized test statistic, and its corresponding value with S, enter

```
> cfat <- cbind(HWFAT, TANFAT, SKFAT)
> RK <- t(apply(cfat,1, rank))
> OBSandRK <- cbind(cfat, RK)
> OBSandRK[1:5,]
      HWFAT TANFAT SKFAT HWFAT TANFAT SKFAT
[1,] 10.71   11.9  9.80     2     3     1
[2,]  8.53   10.0 10.56     1     2     3
[3,]  6.78    8.3  8.43     1     2     3
[4,]  9.32    8.2 11.77     2     1     3
[5,] 41.89   41.6 41.09     3     2     1
> Rj <- apply(RK,2, sum)
> b <- length(HWFAT)
> k <- dim(cfat)[2]
> S <- (12/(b*k*(k+1)))*sum(Rj^2)-3*b*(k+1)
> S
[1] 22.48718
> pval <- 1-pchisq(S, k-1)
> pval
[1] 1.309095e-05
```

To find the standardized test statistic and its corresponding φ -value with the function `friedman.test()`, enter

```
> friedman.test(FAT~GROUP|BLOCK) # R syntax
```

Friedman rank sum test

data: FAT and GROUP and BLOCK

Friedman chi-squared = 22.4872, df = 2, p-value = 1.309e-05

Since the null hypothesis of equal medians is soundly rejected, at least two of the three body fat measuring techniques have different medians. Using (10.20) with an $\alpha = 0.20$, all three of the body fat measuring techniques are declared to be significantly different from each other since $ZR_{12\text{obs}} = 4.72 > Z_{1-\alpha/(k(k-1))} = 1.83$, $ZR_{13\text{obs}} = 2.00 > Z_{1-\alpha/(k(k-1))} = 1.83$, and $ZR_{23\text{obs}} = 2.72 > Z_{1-\alpha/(k(k-1))} = 1.83$. In this case, the probability that all the statements are correct is $1 - \alpha = 0.8$. Since HWFAT is the accepted standard for measuring body fat,

neither of the other two methods is an acceptable substitute for measuring body fat for high school wrestlers.

Code to compute the multiple comparisons according to (10.20) is

```
> alpha <- 0.20
> ZR12 <- abs(Rj[1]-Rj[2])/sqrt(b*k*(k+1)/6)
> ZR13 <- abs(Rj[1]-Rj[3])/sqrt(b*k*(k+1)/6)
> ZR23 <- abs(Rj[2]-Rj[3])/sqrt(b*k*(k+1)/6)
> CV <- round(qnorm(1- alpha/(k*(k-1))),2)
> ZRij <- round(c(ZR12, ZR13, ZR23),2)
> names(ZRij) <- c("ZR12", "ZR13", "ZR23")
> ZRij
ZR12 ZR13 ZR23
4.72 2.00 2.72
> CV
[1] 1.83
> which(ZRij > CV)
ZR12 ZR13 ZR23
  1    2    3
```



10.7 Goodness-of-Fit Tests

Many statistical procedures require knowledge of the population from which the sample is taken. For example, using Student's t -distribution for testing a hypothesis or constructing a confidence interval for μ assumes that the parent population is normal. In this section, **goodness-of-fit** (GOF) procedures are presented that will help to identify the distribution of the population from which the sample is drawn. The null hypothesis in a goodness-of-fit test is a statement about the form of the cumulative distribution. When all the parameters in the null hypothesis are specified, the hypothesis is called simple. Recall that in the event the null hypothesis does not completely specify all of the parameters of the distribution, the hypothesis is said to be composite. Goodness-of-fit tests are typically used when the form of the population is in question. In contrast to most of the statistical procedures discussed so far, where the goal has been to reject the null hypothesis, in a GOF test one hopes to retain the null hypothesis. Two general approaches, one designed primarily for discrete distributions (chi-square goodness-of-fit) and one designed primarily for continuous distributions (Kolmogorov-Smirnov), are presented.

10.7.1 The Chi-Square Goodness-of-Fit Test

Given a single random sample of size n from an unknown population F_X , one may wish to test the hypothesis that F_X has some known distribution $F_0(x)$ for all x . For example, using the data frame `Soccer` from Example 4.4 on page 122, is it reasonable to assume the number of goals scored during regulation time for the 232 soccer matches has a Poisson distribution with $\lambda = 2.5$? Although the problem was previously analyzed, it will be considered again shortly in the context of the chi-square goodness-of-fit test. The chi-square goodness-of-fit test is based on a normalized statistic that examines the vertical deviations between what is observed and what is expected when H_0 is true in k mutually exclusive

categories. At times, such as in surveys of brand preferences, where the categories/groups would be the brand names, the sample will lend itself to being divided into k mutually exclusive categories. Other times, the categories/groupings will be more arbitrary. Before applying the chi-square goodness-of-fit test, the data must be grouped according to some scheme to form k mutually exclusive categories. When the null hypothesis completely specifies the population, the probability that a random observation will fall into each of the chosen or fixed categories can be computed. Once the probabilities for a data point to fall into each of the chosen or fixed categories is computed, multiplying the probabilities by n produces the expected counts for each category under the null distribution. If the null hypothesis is true, the differences between the counts observed in the k categories and the counts expected in the k categories should be small. The test criterion for testing $H_0 : F_X(x) = F_0(x)$ for all x against the alternative $H_1 : F_X(x) \neq F_0(x)$ for some x when the null hypothesis is completely specified is

$$\chi_{\text{obs}}^2 = \sum_{i=1}^k \frac{(O_k - E_k)^2}{E_k}, \quad (10.21)$$

where χ_{obs}^2 is the sum of the squared deviations between what is observed (O_k) and what is expected (E_k) in each of the k categories divided by what is expected in each of the k categories. Large values of χ_{obs}^2 occur when the observed data are inconsistent with the null hypothesis and thus lead to rejection of the null hypothesis. The exact distribution of χ_{obs}^2 is very complicated; however, for large n , provided all expected categories are at least 5, χ_{obs}^2 is distributed approximately χ^2 with $k - 1$ degrees of freedom. When the null hypothesis is composite, that is, not all of the parameters are specified, the degrees of freedom for the random variable χ_{obs}^2 are reduced by one for each parameter that must be estimated.

Example 10.12 ▷ *Soccer Goodness-of-Fit* ◁ Test the hypothesis that the number of goals scored during regulation time for the 232 soccer matches stored in the data frame **Soccer** has a Poisson **cdf** with $\lambda = 2.5$ with the chi-square goodness-of-fit test and an α level of 0.05. Produce a histogram showing the number of observed goals scored during regulation time and superimpose on the histogram the number of goals that are expected to be made when the distribution of goals follows a Poisson distribution with $\lambda = 2.5$.

Solution: Since the number of categories for a Poisson distribution is theoretically infinite, a table is first constructed of the observed number of goals to get an idea of reasonable categories.

```
> attach(Soccer)
> table(Goals)
Goals
 0  1  2  3  4  5  6  7  8
19 49 60 47 32 18  3  3  1
```

Based on the table, a decision is made to create categories for 0, 1, 2, 3, 4, 5, and 6 or more goals. Under the null hypothesis that $F_0(x)$ is a Poisson distribution with $\lambda = 2.5$, the probabilities of scoring 0, 1, 2, 3, 4, 5, and 6 or more goals are computed with **S** as follows:

```
> PX <- c(dpois(0:5,2.5),1-ppois(5,2.5))
> PX
[1] 0.08208500 0.20521250 0.25651562 0.21376302 0.13360189 0.06680094
[7] 0.04202104
```

Since there were a total of $n = 232$ soccer games, the expected number of goals for the six categories is simply $232 \times \text{PX}$:

```
> EX <- 232*PX
> OB <- c(19,49,60,47,32,18,7)
> ans <- cbind(PX, EX, OB)
> row.names(ans) <- c(" X=0", " X=1", " X=2", " X=3", " X=4", " X=5",
+ "X>=6")
> ans
```

	PX	EX	OB
X=0	0.08208500	19.04372	19
X=1	0.20521250	47.60930	49
X=2	0.25651562	59.51162	60
X=3	0.21376302	49.59302	47
X=4	0.13360189	30.99564	32
X=5	0.06680094	15.49782	18
X>=6	0.04202104	9.74888	7

Step 1: **Hypotheses** — The null and alternative hypotheses for using the chi-square goodness-of-fit test to test the hypothesis that the number of goals scored during regulation time for the 232 soccer matches stored in the data frame **Soccer** has a Poisson **cdf** with $\lambda = 2.5$ are

$$H_0 : F_X(x) = F_0(x) \sim \text{Pois}(\lambda = 2.5) \text{ for all } x \text{ versus}$$

$$H_1 : F_X(x) \neq F_0(x) \text{ for some } x.$$

Step 2: **Test Statistic** — The test statistic chosen is χ_{obs}^2 .

Step 3: **Rejection Region Calculations** — Reject if $\chi_{\text{obs}}^2 > \chi_{1-\alpha; k-1}^2$. χ_{obs}^2 is computed with (10.21):

```
> chi.obs <- sum((OB-EX)^2/EX)
> chi.obs
[1] 1.391940
```

$$1.391940 = \chi_{\text{obs}}^2 \stackrel{?}{>} \chi_{.95;6}^2 = 12.59.$$

Step 4: **Statistical Conclusion** — The φ -value is 0.9663469.

```
> p.val <- 1-pchisq(chi.obs,7-1)
> p.val
[1] 0.9663469
```

- I. Since $\chi_{\text{obs}}^2 = 1.391940$ is not greater than $\chi_{.95;6}^2 = 12.59$, fail to reject H_0 .
- II. Since the φ -value = 0.9663469 is greater than 0.05, fail to reject H_0 .

Fail to reject H_0 .

Step 5: **English Conclusion** — There is no evidence to suggest that the true **cdf** does not equal the Poisson distribution with $\lambda = 2.5$ for at least one x .

R and S-PLUS use different functions to perform a chi-square goodness-of-fit test. The R function to implement a chi-square goodness-of-fit test is `chisq.test()`, and the S-PLUS function is `chisq.gof()`. Although S-PLUS has the function `chisq.test()`, it will not accept all of the arguments that the R function `chisq.test()` does. The reader should refer to Table A.6 on page 663 or the respective help files for more information on each function.

Code and output for R:

```
> chisq.test(x=OB, p=PX)
```

```
Chi-squared test for given probabilities
```

```
data: OB
```

```
X-squared = 1.3919, df = 6, p-value = 0.9663
```

Code and output for S-PLUS follow. The argument `cut.points` specifies the categories with an open left interval and closed right interval, that is, `(lower, upper]`. By specifying `right=TRUE`, the categories become left closed and right open.

```
> X2obs <- chisq.gof(Goals, cut.points=c(-1,0,1,2,3,4,5, Inf),
+ distribution="poisson", lambda=2.5)
> X2obs
```

```
Chi-square Goodness of Fit Test
```

```
data: Goals
```

```
Chi-square = 1.3919, df = 6, p-value = 0.9663
```

```
alternative hypothesis:
```

```
True cdf does not equal the poisson Distn. for at least
one sample point.
```

```
> ans <- as.matrix(cbind(X2obs$counts, X2obs$expected))
> rows <- c(" X=0"," X=1"," X=2"," X=3"," X=4"," X=5"," X>=6")
> cols <- c("Observed","Expected")
> dimnames(ans) <- list(rows, cols)
> ans
```

	Observed	Expected
X=0	19	19.0437
X=1	49	47.6093
X=2	60	59.5116
X=3	47	49.5930
X=4	32	30.9956
X=5	18	15.4978
X>=6	7	9.7489

The S code used to create a histogram with superimposed expected goals is

```
> hist(Goals,breaks=c((-0.5):(8.5)), col=13, ylab="", freq=TRUE, main="")
> x <- 0:8
> fx <- (dpois(0:8, lambda=2.5))*232
> lines(x, fx, type="h")
> lines(x, fx, type="p", pch=16)
```

Note that the histogram does not reflect the category ≥ 6 , but rather depicts the observed categories of 6, 7, and 8.

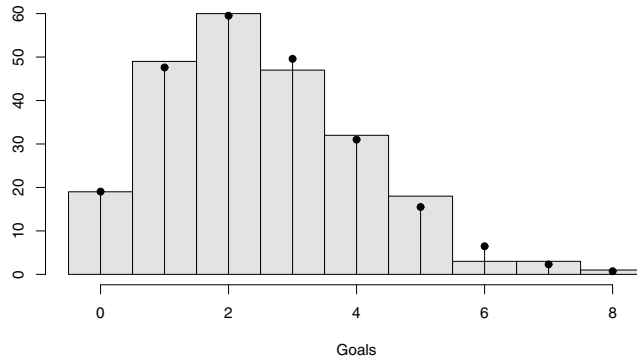


FIGURE 10.11: Histogram of observed goals for **Soccer** with a superimposed Poisson distribution with $\lambda = 2.5$ (vertical lines) ■

Although the chi-square goodness-of-fit test is primarily designed for discrete distributions, it can also be used with a continuous distribution if appropriate categories are defined.

Example 10.13 ▷ *Goodness-of-Fit for SAT Scores* ◁ Use the chi-square goodness-of-fit test with $\alpha = 0.05$ to test the hypothesis that the SAT scores stored in the data frame **Grades** have a normal **cdf**. Use categories, $(-\infty, \mu - 2\sigma]$, $(\mu - 2\sigma, \mu - \sigma]$, $(\mu - \sigma, \mu]$, $(\mu, \mu + \sigma]$, $(\mu + \sigma, \mu + 2\sigma]$, and $(\mu + 2\sigma, \infty]$. Produce a histogram using the categories specified and superimpose on the histogram the expected number of SAT scores in each category when $F_0(x) \sim N(\mu = \bar{x}, \sigma = s)$.

Solution: The test follows:

Step 1: **Hypotheses** — The null and alternative hypotheses for using the chi-square goodness-of-fit test to test the hypothesis that the SAT scores stored in the data frame **Grades** have a Normal **cdf** are

$$H_0 : F_X(x) = F_0(x) \sim N(\mu = \bar{x}, \sigma = s) \text{ for all } x \text{ versus}$$

$$H_1 : F_X(x) \neq F_0(x) \text{ for some } x.$$

Step 2: **Test Statistic** — Since the mean and standard deviation are unknown, the first step is to estimate the unknown parameters μ and σ using $\bar{x} = 11334.65$ and $s = 145.61$:

```
> attach(grades)
> mu <- mean(sat)
```

```
> sig <- sd(sat) # stdev(sat) for S-PLUS
> c(mu, sig)
[1] 1134.6500 145.6087
```

Because a normal distribution is continuous, it is necessary to create categories that include all the data. The categories $\mu - 3\sigma$ to $\mu - 2\sigma$, \dots , $\mu + 2\sigma$ to $\mu + 3\sigma$ are 697.82 to 843.43, 843.43 to 989.04, 989.04 to 1134.65, 1123.65 to 1280.26, 1280.26 to 1425.87, and 1425.87 to 1571.48. These particular categories include all of the observed SAT scores; however, the probabilities actually computed for the largest and smallest categories will be all of the area to the right and left, respectively, of $\bar{x} \pm 2s$. This is done so that the total area under the distribution in the null hypothesis is one.

```
> bin <- seq(mu-3*sig, mu+3*sig, sig)
> bin
[1] 697.8240 843.4326 989.0413 1134.6500 1280.2587 1425.8674
[7] 1571.4760
> table(cut(sat, breaks=bin))

          (698,843]          (843,989]          (989,1.13e+03]
              4                  27                  65
(1.13e+03,1.28e+03] (1.28e+03,1.43e+03] (1.43e+03,1.57e+03]
              80                  21                  3
> OB <- hist(sat, breaks=bin, plot=F)$counts
> PR <- c(pnorm(-2), pnorm(-1:2)- pnorm(-2:1), 1-pnorm(2))
> EX <- 200*PR
> ans <- cbind(PR, EX, OB)
> ans
      PR      EX OB
[1,] 0.02275013 4.550026 4
[2,] 0.13590512 27.181024 27
[3,] 0.34134475 68.268949 65
[4,] 0.34134475 68.268949 80
[5,] 0.13590512 27.181024 21
[6,] 0.02275013 4.550026 3
```

Step 3: **Rejection Region Calculations** — Reject if $\chi_{\text{obs}}^2 > \chi_{1-\alpha; k-p-1}^2$.

Now that the expected and observed counts for each of the categories are computed, the χ_{obs}^2 value can be computed according to (10.21) as 4.173654:

```
> chi.obs <- sum((OB-EX)^2/EX)
> chi.obs
[1] 4.173654
```

Step 4: **Statistical Conclusion** — In this problem, two parameters were estimated, and as a consequence, the degrees of freedom are computed as $6 - 2 - 1 = 3$. The ϕ -value is 0.2433129.

```
> p.val <- 1-pchisq(chi.obs,6-2-1)
> p.val
[1] 0.2433129
```

- I. Since $\chi_{\text{obs}}^2 = 4.173654$ is not greater than $\chi_{.95;3}^2 = 7.81$, fail to reject H_0 .
- II. Since the ϕ -value = 0.2433129 is greater than 0.05, fail to reject H_0 .

Fail to reject H_0 .

Step 5: **English Conclusion** — There is no evidence to suggest that the true **cdf** of SAT scores is not a normal distribution.

If one uses the R function `chisq.test()`, the degrees of freedom and the subsequent ϕ -value will be incorrect, as illustrated next:

```
> chisq.test(x=OB, p=PR)
```

```
Chi-squared test for given probabilities
```

```
data: OB X-squared = 4.1737, df = 5, p-value = 0.5247
```

The S-PLUS function `chisq.gof()` computes the degrees of freedom and the corresponding ϕ -value correctly, provided the argument `n.param.est=` is correctly specified:

```
> chisq.gof(sat, cut.points=c(-Inf, bin[2:6], Inf),
+ distribution="normal", mean=mu, sd=sig, n.param.est=2)
```

Warning messages:

```
Expected counts < 5. Chi-squared approximation may not
be appropriate. in: chisq.gof(sat, cut.points = c( -
Inf, bin[2:6], Inf), ....
```

```
Chi-square Goodness of Fit Test
```

```
data: sat Chi-square = 4.1737, df = 3,
p-value = 0.2433
```

```
alternative hypothesis:
```

```
True cdf does not equal the normal Distn. for at least
one sample point.
```

Since it is not feasible to produce a histogram that extends from $-\infty$ to ∞ , a histogram is created where the categories will simply cover the range of observed values. In this problem, the range of the SAT scores is 720 to 1550. The histogram with categories $(\mu - 3\sigma, \mu - 2\sigma]$, $(\mu - 2\sigma, \mu - \sigma]$, $(\mu - \sigma, \mu]$, $(\mu + \sigma, \mu + 2\sigma]$, and $(\mu + 2\sigma, \mu + 3\sigma]$, superimposed with the expected number of SAT scores for the categories $(-\infty, \mu - 2\sigma]$, $(\mu - 2\sigma, \mu - \sigma]$, $(\mu - \sigma, \mu]$, $(\mu, \mu + \sigma]$, $(\mu + \sigma, \mu + 2\sigma]$, and $(\mu + 2\sigma, \infty]$ is computed with the code given next and depicted in Figure 10.12 on the next page.

```
> hist(sat, breaks=bin, col=13, ylab="", freq=TRUE, main="")
> x <- bin[2:7]-sig/2
> fx <- PR*200
> lines(x, fx, type="h")
> lines(x, fx, type="p", pch=16)
```

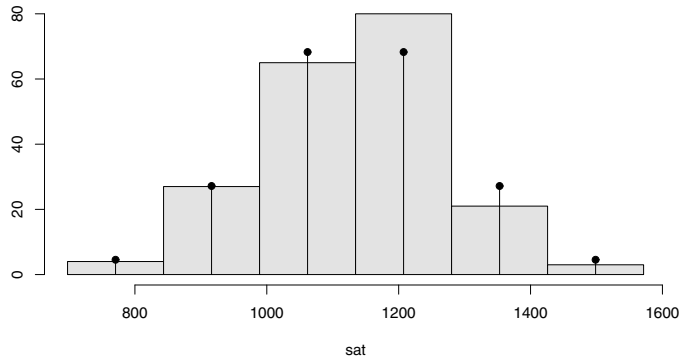



FIGURE 10.12: Histogram of SAT scores in **Grades** superimposed with the expected number of SAT scores for the categories $(-\infty, \mu - 2\sigma]$, $(\mu - 2\sigma, \mu - \sigma]$, $(\mu - \sigma, \mu]$, $(\mu, \mu + \sigma]$, $(\mu + \sigma, \mu + 2\sigma]$, and $(\mu + 2\sigma, \infty]$ (vertical lines) ■

10.7.2 Kolmogorov-Smirnov Goodness-of-Fit Test

In Section 10.7.1, the chi-square goodness-of-fit test worked by measuring the vertical distance between what was observed in a particular category and what was expected in that same category under the null hypothesis for each of the k categories. In contrast to the chi-square goodness-of-fit test, the Kolmogorov-Smirnov goodness-of-fit test uses all n observations and measures vertical deviations between the cumulative distribution function (**cdf**), $F_0(x)$ (where all parameters are specified), and the empirical cumulative distribution function (**ecdf**), $\hat{F}_n(x)$, for all x . For large n , the deviations between $F_0(x)$ and $\hat{F}_n(x)$ should be small for all values of x . The statistic D_n , called the Kolmogorov-Smirnov one-sample statistic, is defined as

$$D_n = \sup_x \left| \hat{F}_n(x) - F_0(x) \right| \quad (10.22)$$

The statistic D_n does not depend on $F_0(x)$ as long as $F(x)$ is continuous. The derivation of the sampling distribution of D_n is beyond the scope of this text. The curious reader can refer to Gibbons and Chakraborti (2003), page 114, for the derivation of the sampling distribution of D_n . The statistic and sampling distribution of D_n should only be used with simple hypotheses. When the null hypothesis is composite, the critical values for the Kolmogorov-Smirnov test (based on the sampling distribution of D_n) are extremely conservative. The Kolmogorov-Smirnov test can be used to assess normality provided the distribution is completely specified. In a test of normality where the null hypothesis is not completely specified, the statistic D_n can still be used by estimating the unknown parameters of $F_0(x)$ using maximum likelihood ($\hat{F}_0(x)$) and substituting $\hat{F}_0(x)$ for $F_0(x)$ in (10.22). However, this further complicates the sampling distribution of D_n . When testing a composite normal hypothesis with unknown μ and σ , the test that uses $D_n = \sup_x |\hat{F}_n(x) - \hat{F}_0(x)|$ is called Lilliefors's normality test (explained more fully starting on page 458). Lilliefors used simulation to study the sampling distribution of D_n for composite hypotheses and subsequently to publish critical values for using D_n with composite hypotheses. Simulation will be used to show the differences in the distribution of D_n for a simple null hypothesis versus the distribution of D_n with a composite null hypothesis.

Recall that the **ecdf** was defined in (6.2) to be:

$$\hat{F}_n(t) = \sum_{i=1}^n \mathbf{I}\{x_i \leq t\}/n$$

An equivalent expression for the **ecdf** is

$$\hat{F}_n(x) = \begin{cases} 0 & x < X_{(1)} \\ \frac{i}{n} & X_{(i)} \leq x \leq X_{(i+1)} \\ 1 & x > X_{(n)} \end{cases} \quad (10.23)$$

which will prove useful in computing D_n . When all n observations are distinct, D_n can be computed as

$$D_n = \max_{i=1, \dots, n} M_i \quad (10.24)$$

where

$$M_i = \max \left\{ \left| \hat{F}_n(X_{(i)}) - F_0(X_{(i)}) \right|, \left| F_0(X_{(i)}) - \hat{F}_n(X_{(i-1)}) \right| \right\} \quad (10.25)$$

Since $\hat{F}_n(X_{(i)}) = \frac{i}{n}$ and $\hat{F}_n(X_{(i-1)}) = \frac{i-1}{n}$, (10.25) can be expressed as

$$M_i = \max \left\{ \left| \frac{i}{n} - F_0(X_{(i)}) \right| = D_i^+, \left| F_0(X_{(i)}) - \frac{i-1}{n} \right| = D_i^- \right\} \quad (10.26)$$

Stated formally, the null and alternative hypotheses for the Kolmogorov-Smirnov test for goodness-of-fit are

$$H_0 : F(x) = F_0(x) \text{ for all } x \text{ versus } H_1 : F(x) \neq F_0(x) \text{ for some } x. \quad (10.27)$$

The null hypothesis is rejected when $D_n > D_{n;1-\alpha}$ or when the test's \wp -value is less than the largest acceptable α value. Since **S** will compute the \wp -value for the Kolmogorov-Smirnov test, critical values for various n and α are not presented. **R** uses the function `ks.test(x, y, ...)`, where **x** is a numeric vector of observations and **y** is either a numeric vector of data values or a character string naming a distribution function. **S-PLUS** uses the function `ks.gof(x, distribution = "normal", ...)`, where **x** is a numeric vector of observations. The examples will illustrate the use of both functions.

Example 10.14 ▷ *Kolmogorov-Smirnov GOF Test* ◁ Test whether the observations 5, 6, 7, 8, and 9 are from a normal distribution with $\mu = 6.5$ and $\sigma = \sqrt{2}$. That is, the hypothesized distribution is $F_0(x) \sim N(6.5, \sqrt{2})$.

Solution: Since $F_0(x) \sim N(6.5, \sqrt{2})$, it follows that

$$F_0(X_{(i)}) = P(Y \leq X_{(i)}) = P\left(\frac{Y - 6.5}{\sqrt{2}} \leq \frac{X_{(i)} - 6.5}{\sqrt{2}}\right) = P\left(Z \leq \frac{X_{(i)} - 6.5}{\sqrt{2}}\right).$$

To compute $F_0(X_{(i)})$ with **S**, key in

```
> x <- 5:9
> mu <- 6.5
> sig <- sqrt(2)
> x <- sort(x)
> n <- length(x)
> FoX <- pnorm(x, mean=mu, sd=sig)
> FoX
[1] 0.14442 0.36184 0.63816 0.85558 0.96145
```

The quantities $\hat{F}_n(X_{(i)}) = \frac{i}{n}$, $\hat{F}_n(X_{(i-1)}) = \frac{i-1}{n}$, D_i^+ , D_i^- , and M_i are computed and stored in the S variables `FnX`, `Fn1X`, `Dp`, `Dm`, and `Mi`, respectively. The Komolgorov-Smirnov statistic $D_n = \max_{i=1, \dots, n} M_i$ is 0.25558.

```
> FnX <- seq(1:n)/n
> Fn1X <- (seq(1:n)-1)/n
> DP <- (FnX - FoX)
> DM <- FoX - Fn1X
> Dp <- abs(DP)
> Dm <- abs(DM)
> EXP <- cbind(x, FnX, Fn1X, FoX, Dp, Dm)
> Mi <-apply(EXP[, c(5,6)],1, max)
> TOT <- cbind(EXP, Mi)
> TOT
      x FnX Fn1X      FoX      Dp      Dm      Mi
[1,] 5 0.2  0.0 0.14442 0.055578 0.14442 0.14442
[2,] 6 0.4  0.2 0.36184 0.038163 0.16184 0.16184
[3,] 7 0.6  0.4 0.63816 0.038163 0.23816 0.23816
[4,] 8 0.8  0.6 0.85558 0.055578 0.25558 0.25558
[5,] 9 1.0  0.8 0.96145 0.038550 0.16145 0.16145
> Dn <- max(Mi)
> Dn
[1] 0.25558
```

Table 10.20: Calculating D_n

i	$X_{(i)}$	$\frac{i}{n} - F_0(X_{(i)})$	$F_0(X_{(i)}) - \frac{i-1}{n}$	D^+	D^-	M_i
1	5	$\frac{1}{5} - 0.14442$	$0.14442 - 0$	0.055578	0.14442	0.14442
2	6	$\frac{2}{5} - 0.36184$	$0.36184 - \frac{1}{5}$	0.038163	0.16184	0.16184
3	7	$\frac{3}{5} - 0.63816$	$0.63816 - \frac{2}{5}$	0.038163	0.23816	0.23816
4	8	$\frac{4}{5} - 0.85558$	$0.85558 - \frac{3}{5}$	0.055578	0.25558	0.25558
5	9	$\frac{5}{5} - 0.96145$	$0.96145 - \frac{4}{5}$	0.038550	0.16145	0.16145
					$D_n =$	0.25558

The computation of the Komolgorov-Smirnov statistic D_n and its φ -value with R and S-PLUS (given that the previous S code has been entered) follow.

R code:

```
> ks.test(x, y="pnorm", mean=mu, sd=sig)
```

One-sample Kolmogorov-Smirnov test

data: x

D = 0.2556, p-value = 0.827

alternative hypothesis: two.sided

S-PLUS code:

```
> ks.gof(x, dist="normal", mean=mu, sd=sig)
```

```
One-sample Kolmogorov-Smirnov Test
Hypothesized distribution = normal
```

```
data: x
```

```
ks = 0.2556, p-value = 0.8269
```

```
alternative hypothesis:
```

```
True cdf is not the normal distn. with the specified parameters
```

The Komolgorov-Smirnov statistic is labeled D in R and ks in S-PLUS. Both R and S-PLUS return the value $D_n = 0.2556$ with a corresponding ϕ -value of 0.827, which provides no evidence to reject the null hypothesis that $F_0(x) \sim N(6.5, \sqrt{2})$. Figure 10.13 provides a graphical illustration of the vertical deviations used to compute the statistic D_n for this problem.

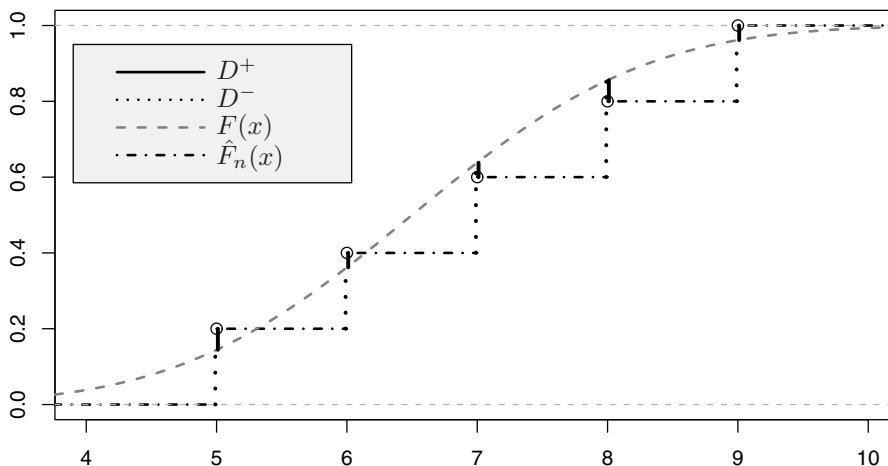


FIGURE 10.13: Graphical illustration of the vertical deviations used to compute the statistic D_n for Example 10.14 on page 455 ■

In Example 10.14 on page 455, the statistic $D_n = 0.2556$ was computed, and both R and S-PLUS returned a ϕ -value of 0.827. To visualize the sampling distribution of D_n and to find simulated critical values, one can use R code similar to the following:

```
ksdist <- function (n = 10, sims = 10000, alpha = 0.05){
  Dn <- replicate(sims, ks.test(rnorm(n), pnorm)$statistic)
  cv <- quantile(Dn, 1 - alpha)
  plot(density(Dn), col = "blue", lwd = 2, main = "",
       xlab = paste("Simulated critical value =", round(cv,3),
                    "for n =", n, "when the alpha value =", alpha))
  title(main=list(expression(paste("Simulated Sampling Distribution of ",
                                  D[n]))))
}
```

The graph from running `ksdist(n=5, sims=10000, alpha=0.05)` is shown in Figure 10.14. This simulation indicates a value of 0.567 or greater would be required to reject the null hypothesis in Example 10.14 on page 455 at the $\alpha = 0.05$ level. The simulated ϕ -value for the value $D_n = 0.257$ in Figure 10.14 is 0.827, the same value as reported by both R and S-PLUS using `ks.test()` and `ks.gof()`, respectively.

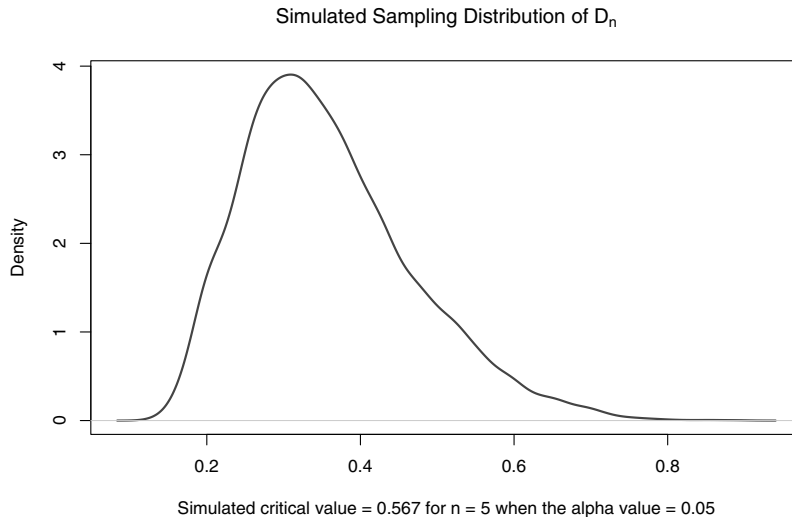


FIGURE 10.14: Graphical illustration of `ksdist(n=5, sims=10000, alpha=0.05)`

Lilliefors's Test of Normality

Expanding on the simulation for the sampling distribution for D_n (only works for R), consider what happens when the null hypothesis changes from simple to composite using the code for the function `ksLdist()`. Note that the D_n values stored in `D_n[i]` are for a simple null hypothesis of normality while the D_n values stored in `DnL[i]` are for a composite hypothesis of normality. The critical values reported by Lilliefors (1967) were based on simulations using 1000 or more samples using logic similar to the R code for `ksLdist()`:

```
ksLdist <- function (n = 10, sims = 1000, alpha = 0.05)
{
  Dn <- c()
  DnL <- c()
  for (i in 1:sims) {
    x <- rnorm(n)
    mu <- mean(x)
    sig <- sd(x)
    Dn[i] <- ks.test(x, pnorm)$statistic
    DnL[i] <- ks.test(x, pnorm, mean = mu, sd = sig)$statistic
  }
  ys <- range(density(DnL)$y)
  xs <- range(density(Dn)$x)
```

```

cv <- quantile(Dn, 1 - alpha)
cvp <- quantile(DnL, 1 - alpha)
plot(density(Dn, bw=0.02), col="blue", lwd=2, ylim=ys, xlim=xs,
     main = "", , xlab="", sub = paste("Simulated critical value =",
     round(cv, 3), "(simple hypothesis) and ", round(cvp, 3),
     "(composite hypothesis)\n for n=", n,"when the alpha value =",
     alpha))
title(main=list(expression(paste("Simulated Sampling Distribution of ",
D[n])))))
lines(density(DnL, bw = 0.02), col = "red", lwd = 2, lty = 2)
legend(mean(xs), max(ys), legend = c("Simple Hypothesis",
"Composite Hypothesis"), col = c("blue", "red"), xjust = 0,
text.col = c("black", "black"), lty = c(1, 2), bg = "gray95",
cex = 1, lwd = 2)
box()
abline(h = 0)
}

```

The function `ksLdist()` allows the user to choose the number of samples with the argument `sims=` and easily to verify the results given by Lilliefors (1967). Dallal and Wilkinson (1986) duplicated the work by Lilliefors (1967) using much larger samples as well as deriving an analytic approximation for the upper tail φ -values for $D_n = \sup_x |\hat{F}_n(x) - \hat{F}_0(x)|$. For φ -values less than 0.100 and sample sizes ranging from 5 to 100, the Dallal-Wilkinson approximation is

$$\widehat{\varphi\text{-value}} = \exp(-7.01256 \cdot D_n^2 \cdot (n + 2.78019) + 2.99587 \cdot D_n \cdot \sqrt{n + 2.78019} - 0.122119 + 0.974598/\sqrt{n} + 1.67997/n) \quad (10.28)$$

The estimated densities from running `ksLdist(sims=10000, n=10)` are shown in Figure 10.15 on the following page, which highlights how much less variability is present in the sampling distribution of D_n when the null hypothesis is composite. To correctly test a composite hypothesis of normality, one should use the R function `lillie.test()` available in the R package `nortest`. That is, one should not use the R function `ks.test()`. However, both simple and composite hypotheses of normality can be tested using the S-PLUS function `ks.gof()`. The R function `lillie.test()` from the `nortest` package produces virtually the same result as the S-PLUS function `ks.gof()`, with the exception that the φ -value is not set to 0.5 when the Dallal-Wilkinson approximation yields a φ -value greater than 0.1 when testing a composite hypothesis of normality.

Example 10.15 ▷ *Long Distance Phone Calls* ◁ Calculate the φ -value and state the English conclusion for testing whether the times spent on long distance phone calls (`call.time`) in the data frame `Phone` have a normal distribution using the R function `lillie.test` from the `nortest` package as well as using the S-PLUS function `ks.gof()`. Verify the reported φ -values using (10.28).

Solution: The results after attaching the data frame `Phone` are presented first for R, and then for S-PLUS. Note that R labels the statistic D_n with `D` and S-PLUS with `ks`. Both functions return $D_n = 0.1910237$ with a φ -value = 0.0291.

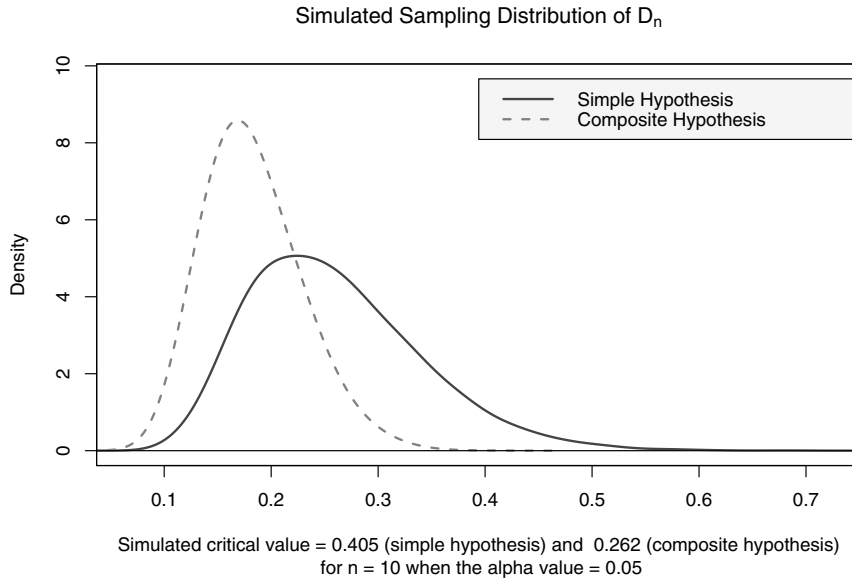


FIGURE 10.15: Estimated densities for simple and composite hypotheses from running `ksLdist(sims=10000, n=10)`

For R:

```
> attach(Phone)
> library(nortest)
> lillie.test(call.time)
```

Lilliefors (Kolmogorov-Smirnov) normality test

```
data: call.time
D = 0.191, p-value = 0.0291
```

For S-PLUS:

```
> attach(Phone)
> ks.gof(call.time, distribution="normal")
```

One sample Kolmogorov-Smirnov Test of Composite Normality

```
data: call.time
ks = 0.191, p-value = 0.0291
alternative hypothesis:
  True cdf is not the normal distn. with estimated parameters
sample estimates:
  mean of x standard deviation of x
  3.686957          3.623698
```

To compute the φ -value using (10.28), a small function `DWA` is written that returns an estimated φ -value of 0.0291:

```

> DWA <- function(Dn = .3, n =10)
+ {
+ p.value<- exp(-7.01256*Dn^2*(n + 2.78019)
+ + 2.99587*Dn*(n + 2.78019)^.5-0.122119
+ + .974598/n^.5+1.67997/n)
+ round(p.value,4)
+ }
> DWA(Dn=0.1910237, n=23)
[1] 0.0291

```

With a p -value of 0.0291, the null hypothesis is rejected. There is evidence that phone call length is not normally distributed. ■

10.7.3 Shapiro-Wilk Normality Test

The Shapiro-Wilk test is appropriate for testing normality. More specifically, the test allows for a composite hypothesis of normality. That is, the parameters of the normal distribution do not need to be specified in the null hypothesis of the test (as they must be for the Lilliefors test). Although the test is known to be conservative, it is useful for testing normality with small samples. The test statistic measures how closely the empirical quantiles of the sample follow the corresponding theoretical quantiles of a normal distribution. This means that small values of the test statistic lead to the rejection of the null hypothesis (that the distribution is normal).

To calculate the test statistic for a random sample of size n , x_1, x_2, \dots, x_n , the sample must be sorted: $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$. The Shapiro-Wilk test statistic takes the form

$$W = \frac{b^2}{nS_u^2}, \quad (10.29)$$

where S_u^2 is the uncorrected sample variance, $b = \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} a_{n-i+1}(x_{(n-i+1)} - x_{(i)})$, and $\lfloor \frac{n}{2} \rfloor$ is the integer part of $\frac{n}{2}$. The coefficients a_{n-i+1} that are calculated automatically by **S** are tabulated in Table 6 of Shapiro and Wilk (1965).

The critical region of the test is given by

$$\mathbb{P}(W \leq K | H_0) = \alpha,$$

where α is the significance level. The critical values K can be found in Shapiro and Wilk (1965, Table 5), but they are not displayed in the **S** output for this test. The vector of weights $\mathbf{a}' = (a_1, \dots, a_n)$, where $a_i = -a_{n-i+1}$, is calculated as

$$\mathbf{a} = \frac{\mathbf{w}'\mathbf{V}^{-1}}{\mathbf{w}'\mathbf{V}^{-1}\mathbf{V}^{-1}\mathbf{w}}, \quad (10.30)$$

where the elements of the vector \mathbf{w} are $w_i = E[x_{(i)}]$ and \mathbf{V} is the covariance matrix of the order statistics $x_{(1)}, x_{(2)}, \dots, x_{(n)}$.

Example 10.16 ▷ *Shapiro-Wilk Normality Test* ◁ Use the Shapiro-Wilk test with the random sample $\{47, 50, 57, 54, 52, 54, 53, 65, 62, 67, 69, 74, 51, 57, 57, 59\}$ to test for normality using $\alpha = 0.05$.

Solution: First, order the data:

$$47 \leq 50 \leq 51 \leq 52 \leq 53 \leq 54 = 54 \leq 57 = 57 = 57 \leq 59 \leq 62 \leq 65 \leq 67 \leq 69 \leq 74.$$

Next, calculate the differences $x_{(n-i+1)} - x_{(i)}$ for $i = 1, 2, \dots, \lfloor \frac{n}{2} \rfloor = 8$:

$$\left| \begin{array}{l} x_{(16)} - x_{(1)} = 74 - 47 = 27 \\ x_{(15)} - x_{(2)} = 69 - 50 = 19 \\ x_{(14)} - x_{(3)} = 67 - 51 = 16 \end{array} \right| \left| \begin{array}{l} x_{(13)} - x_{(4)} = 65 - 52 = 13 \\ x_{(12)} - x_{(5)} = 62 - 53 = 9 \\ x_{(11)} - x_{(6)} = 59 - 54 = 5 \end{array} \right| \left| \begin{array}{l} x_{(10)} - x_{(7)} = 57 - 54 = 3 \\ x_{(9)} - x_{(8)} = 57 - 57 = 0 \end{array} \right|$$

Looking at Table 6 from Shapiro and Wilk (1965) ($n = 16$ and $i = 1, \dots, 8$), one obtains

$$\left| \begin{array}{l} a_{16} = 0.5056 \\ a_{15} = 0.3290 \end{array} \right| \left| \begin{array}{l} a_{14} = 0.2521 \\ a_{13} = 0.1939 \end{array} \right| \left| \begin{array}{l} a_{12} = 0.1447 \\ a_{11} = 0.1005 \end{array} \right| \left| \begin{array}{l} a_{10} = 0.0593 \\ a_9 = 0.0196 \end{array} \right|$$

which means $b = \sum_{i=1}^8 a_{n-i+1}(x_{(n-i+1)} - x_{(i)}) = 28.4392$ and $nS_u^2 = 854$.

The Shapiro-Wilk test statistic value is then

$$W = \frac{b^2}{nS_u^2} = \frac{808.7881}{854} = 0.9471.$$

The critical value K with $\alpha = 0.05$ and $n = 16$ is 0.887. As $W_{obs} = 0.9471 > 0.887$, one fails to reject the null hypothesis of normality.

In S:

```
> x <- c(47, 50, 57, 54, 52, 54, 53, 65, 62, 67, 69, 74, 51, 57, 57, 59)
> shapiro.test(x)
```

Shapiro-Wilk normality test

```
data:  x W = 0.9471, p-value = 0.4445
```



10.8 Categorical Data Analysis

This section provides an overview of two common scenarios where categorical data are generated and explains how each scenario is analyzed. The basic 2×2 contingency table with fixed row totals was introduced in Section 9.9.3, Testing Equality of Proportions with Fisher's exact test. The 2×2 contingency table can be generalized for I rows and J columns and is referred to as an $I \times J$ contingency table. The sampling scheme employed to acquire the information in the table will determine the type of hypothesis that can be tested. Consider the following two scenarios:

SCENARIO ONE: Is there an association between gender and a person's happiness? To investigate whether happiness depends on gender, one might use information collected from the General Social Survey (GSS) (<http://sda.berkeley.edu/GSS>). In each survey, the GSS asks, "Taken all together, how would you say things are these days—would you say that you are very happy, pretty happy, or not too happy?" Respondents to each survey are coded as either male or female. The information in Table 10.21 on the next page shows how a subset of respondents (26-year-olds) were classified with respect to the variables HAPPY and SEX.

SCENARIO TWO: In a double blind randomized drug trial (neither the patient nor the physician evaluating the patient knows the treatment, drug or placebo, the patient

Table 10.21: Twenty-six-year-olds' happiness

SEX	HAPPY		
	Very happy	Pretty happy	Not too happy
Male	110	277	50
Female	163	302	63

receives), 400 male patients with mild dementia were randomly divided into two groups of 200. One group was given a placebo over three months while the second group received an experimental drug for three months. At the end of the three months, the physicians (all psychiatrists) classified the 400 patients into one of three categories: improved, no change, or worse. Are the proportions in the three status categories the same for the two treatments?

Table 10.22: Mild dementia treatment results

Treatment	Status		
	Improve	No Change	Worse
Drug	67	76	57
Placebo	48	73	79

The two scenarios illustrate two different sampling schemes that both result in $I \times J$ contingency tables. In the first scenario, there is a single population (Americans) and individuals are sampled from this single population and classified into one of the IJ cells of the $I \times J$ contingency table based on the $I = 2$ SEX categories and the $J = 3$ HAPPY categories. The format of an $I \times J$ contingency table when sampling from a single population is shown in Table 10.23. The number of observations from the i^{th} row classified into the j^{th} column is denoted by n_{ij} . It follows that the number of observations in the j^{th} column ($1 \leq j \leq J$) is $n_{\bullet j} = n_{1j} + n_{2j} + \dots + n_{Ij}$, while the number of observations in the i^{th} row ($1 \leq i \leq I$) is $n_{i\bullet} = n_{i1} + n_{i2} + \dots + n_{iJ}$.

The true population proportion of individuals in cell (i, j) will be denoted π_{ij} . Under the assumption of independence between row and column variables (SEX and HAPPY in this example), $\pi_{ij} = \pi_{i\bullet} \times \pi_{\bullet j}$, where $\pi_{i\bullet} = \sum_{j=1}^J \pi_{ij}$ and $\pi_{\bullet j} = \sum_{i=1}^I \pi_{ij}$. That is, $\pi_{i\bullet}$ is the proportion of observations in the population classified in category i of the row variable and $\pi_{\bullet j}$ is the proportion of observations in the population classified in category j of the column variable. Since $\pi_{i\bullet}$ and $\pi_{\bullet j}$ are marginal population proportions, it follows that $\hat{\pi}_{i\bullet} = p_{i\bullet} = \frac{n_{i\bullet}}{n}$ and $\hat{\pi}_{\bullet j} = p_{\bullet j} = \frac{n_{\bullet j}}{n}$, where n is the sample size. Under the assumption of independence the expected count for cell (i, j) is $\mu_{ij} = n\pi_{ij} = n\pi_{i\bullet}\pi_{\bullet j}$ and $\hat{\mu}_{ij} = n\hat{\pi}_{ij} = n\hat{\pi}_{i\bullet}\hat{\pi}_{\bullet j} = n\frac{n_{i\bullet}}{n}\frac{n_{\bullet j}}{n} = \frac{n_{i\bullet}n_{\bullet j}}{n}$.

Table 10.23: Contingency table when sampling from a single population

	Col 1	Col 2	...	Col J	Totals
Row 1	n_{11}	n_{12}	...	n_{1J}	$n_{1\bullet}$
Row 2	n_{21}	n_{22}	...	n_{2J}	$n_{2\bullet}$
\vdots	\vdots	\vdots		\vdots	\vdots
Row I	n_{I1}	n_{I2}	...	n_{IJ}	$n_{I\bullet}$
Totals	$n_{\bullet 1}$	$n_{\bullet 2}$...	$n_{\bullet J}$	n

In the second scenario, there are two distinct populations from which samples are taken. The first population is the group of all patients receiving the experimental drug while the second population is the group of all patients receiving a placebo. In this scenario, there are $I = 2$ separate populations and $J = 3$ categories for the $I = 2$ populations. Individuals sampled from the $I = 2$ distinct populations are classified into one of the $J = 3$ status categories. This scenario has fixed row totals whereas the first scenario does not. In the first scenario, only the total sample size, n , is fixed. That is, neither the row nor the column totals are fixed. This is in contrast to scenario two, where the number of patients in each treatment group (row) was fixed. The notation used for an $I \times J$ contingency table when I samples from I distinct populations differs slightly from the notation used in Table 10.23 on the previous page with a contingency table from a single sample.

Since the sample sizes of the I distinct populations are denoted $n_{i\bullet}$, the total for all I samples is denoted by $n_{\bullet\bullet}$ rather than the notation n used for a single sample in Table 10.23 on the preceding page. Table 10.24 shows the general form and notation used for an $I \times J$ contingency table when sampling from I distinct populations. Each observation in each sample is classified into one of J categories. If $n_{i\bullet}$ denotes the number of observations in the i^{th} sample ($1 \leq i \leq I$) and n_{ij} denotes the number of observations from the i^{th} sample classified into the j^{th} category ($1 \leq j \leq J$), it follows that the number of observations in the j^{th} column is $n_{\bullet j} = n_{1j} + n_{2j} + \dots + n_{Ij}$, while the number of observations in the i^{th} row is $n_{i\bullet} = n_{i1} + n_{i2} + \dots + n_{iJ}$.

Table 10.24: General form and notation used for an $I \times J$ contingency table when sampling from I distinct populations

	Category 1	Category 2	...	Category J	Totals
Population 1	n_{11}	n_{12}	...	n_{1J}	$n_{1\bullet}$
Population 2	n_{21}	n_{22}	...	n_{2J}	$n_{2\bullet}$
\vdots	\vdots	\vdots		\vdots	\vdots
Population I	n_{I1}	n_{I2}	...	n_{IJ}	$n_{I\bullet}$
Totals	$n_{\bullet 1}$	$n_{\bullet 2}$...	$n_{\bullet J}$	$n_{\bullet\bullet}$

10.8.1 Test of Independence

Scenario one asks if there is an association between gender and a person's happiness. In Section 3.3.6 on page 86, two events, A and B , were defined as independent when $\mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B)$ or, equivalently, when $\mathbb{P}(A|B) = \mathbb{P}(A)$. If, instead of having a random sample from a single population, an $I \times J$ contingency table consisted of entries from the population, association could be mathematically verified by showing that $\mathbb{P}(n_{ij}) \neq \mathbb{P}(n_{i\bullet}) \times \mathbb{P}(n_{\bullet j})$ for some i and j . If by chance $\mathbb{P}(n_{ij}) = \mathbb{P}(n_{i\bullet}) \times \mathbb{P}(n_{\bullet j})$ for all i and j , then one would conclude there is no association between gender and a person's happiness. That is, the variables gender and happiness would be considered mathematically independent. Since the entire population is not given but rather a sample from a population, the values in the $I \times J$ contingency table can be expected to change from sample to sample. The question is, "By how much can the variables deviate from the mathematical definition of independence and still be considered statistically independent?"

The null and alternative hypotheses to test for independence between row and column variables is written $H_0 : \pi_{ij} = \pi_{i\bullet}\pi_{\bullet j}$ versus $H_1 : \pi_{ij} \neq \pi_{i\bullet}\pi_{\bullet j}$. The test statistic is

$$\chi_{\text{obs}}^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(O_{ij} - E_{ij})^2}{E_{ij}}. \quad (10.31)$$

It compares the observed frequencies in the table with the expected frequencies when H_0 is true. Under the assumption of independence, and when the observations in the cells are sufficiently large (usually greater than 5), $\chi_{\text{obs}}^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \rightsquigarrow \chi_{(I-1)(J-1)}^2$, where $\hat{\mu}_{ij} = \frac{n_{i\bullet}n_{\bullet j}}{n} = E_{ij}$ and $n_{ij} = O_{ij}$. The null hypothesis of independence is rejected when $\chi_{\text{obs}}^2 > \chi_{1-\alpha; (I-1)(J-1)}^2$.

The chi-squared approximation is generally satisfactory if the E_{ij} s ($\hat{\mu}_{ij}$ s) in the test statistic are not too small. Various rules of thumb exist for what might be considered too small. A very conservative rule is to require all E_{ij} s to be 5 or more. This can be accomplished by combining cells with small E_{ij} s and reducing the overall degrees of freedom. At times, it may be permissible to let the E_{ij} of a cell be as low as 0.5.

Test for SCENARIO ONE:

Step 1: **Hypotheses** — $H_0 : \pi_{ij} = \pi_{i\bullet}\pi_{\bullet j}$ (Row and column variables are independent.) versus $H_1 : \pi_{ij} \neq \pi_{i\bullet}\pi_{\bullet j}$ for at least one i, j (Row and column variables are dependent.)

Step 2: **Test Statistic** — The test statistic is

$$\chi_{\text{obs}}^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \rightsquigarrow \chi_{(I-1)(J-1)}^2 = \chi_{(2-1)(3-1)}^2 = \chi_2^2$$

under the assumption of independence. The χ_{obs}^2 value is 4.32.

Step 3: **Rejection Region Calculations** — The rejection region is

$$\chi_{\text{obs}}^2 > \chi_{1-\alpha; (I-1)(J-1)}^2 = \chi_{0.95; 2}^2 = 5.99.$$

Before the statistic $\chi_{\text{obs}}^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ can be computed, the expected counts for each of the ij cells must be calculated. Note that $O_{ij} = n_{ij}$ and $E_{ij} = \frac{n_{i\bullet}n_{\bullet j}}{n}$:

```
> HA <- c(110,277,50,163,302,63)
> HAT <- matrix(data=HA, nrow=2, byrow=TRUE)
> dimnames(HAT) <- list(Gender=c("Male", "Female"),
+ Category=c("Very Happy", "Pretty Happy", "Not To Happy"))
> HAT
      Category
Gender  Very Happy Pretty Happy Not To Happy
Male      110          277          50
Female    163          302          63
> E <- outer(rowSums(HAT), colSums(HAT), "*/")/sum(HAT)
```

```
> E
      Very Happy Pretty Happy Not To Happy
Male   123.6280      262.2    51.17202
Female 149.3720      316.8    61.82798
```

$$\chi_{\text{obs}}^2 = \frac{(110 - 123.6280)^2}{123.6280} + \frac{(277 - 262.2)^2}{262.2} + \dots + \frac{(63 - 61.83)^2}{61.83} = 4.32$$

The value of the test statistic is $\chi_{\text{obs}}^2 = 4.32$. This can be done with code by entering

```
> chi.obs <- sum((HAT-E)^2/E )
> chi.obs
[1] 4.321482
```

$$4.32 = \chi_{\text{obs}}^2 \stackrel{?}{>} \chi_{.95,2}^2 = 5.99.$$

Step 4: **Statistical Conclusion** — The φ -value is 0.115.

```
> p.val <- 1-pchisq(chi.obs,2)
> p.val
[1] 0.1152397
```

- I. From the rejection region, since $\chi_{\text{obs}}^2 = 4.32 < \chi_{0.95;2}^2 = 5.99$, fail to reject the null hypothesis of independence.
- II. Since the φ -value = .115 is greater than 0.05, fail to reject the null hypothesis of independence.

Fail to reject H_0 .

Step 5: **English Conclusion** — There is not sufficient evidence to suggest the variables gender and happiness are statistically dependent.

Both R and S-PLUS have the function `chisq.test()`, which can be used to test the null hypothesis of independence by computing the observed test statistic and its corresponding φ -value:

```
> chisq.test(HAT)
```

Pearson's Chi-squared test

```
data: HAT
```

```
X-squared = 4.3215, df = 2, p-value = 0.1152
```



10.8.2 Test of Homogeneity

The question of interest in scenario two is whether the proportions in each of the $j = 3$ categories for the $i = 2$ populations are equivalent. Specifically, is $\pi_{1j} = \pi_{2j}$ for all j ? This question is answered with a test of homogeneity. In general, the null hypothesis for a test of homogeneity with $i = I$ populations is written

$$H_0 : \pi_{1j} = \pi_{2j} = \dots = \pi_{Ij} \text{ for all } j \text{ versus } H_1 : \pi_{ij} \neq \pi_{i+1,j} \text{ for some } (i, j). \quad (10.32)$$

Expressed in words, the null hypothesis is that the I populations are homogeneous with respect to the J categories versus the I populations are not homogeneous with respect to the J categories. An equivalent interpretation is that for each population $j = 1, 2, \dots, J$, the proportion of people in the j^{th} category is the same. When H_0 is true, $\pi_{1j} = \pi_{2j} = \dots = \pi_{Ij}$ for all j . Under the null hypothesis, $\mu_{ij} = n_{i\bullet}\pi_{ij}$, $\hat{\pi}_{ij} = p_{ij} = \frac{n_{ij}}{n_{i\bullet}}$, and $\hat{\mu}_{ij} = \frac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}} = E_{ij}$. When H_0 is true, all the probabilities in the j^{th} column are equal, and a pooled estimate of π_{ij} is obtained by adding all the frequencies in the j^{th} column ($n_{\bullet j}$) and dividing the total by $n_{\bullet\bullet}$. The statistic used in this type of problem has the same form as the one used for the test of independence in (10.31). Substituting the homogeneity expressions for O_{ij} and E_{ij} , the statistic is expressed as

$$\chi_{\text{obs}}^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - n_{i\bullet}n_{\bullet j}/n_{\bullet\bullet})^2}{n_{i\bullet}n_{\bullet j}/n_{\bullet\bullet}} \dot{\sim} \chi_{(I-1)(J-1)}^2.$$

The null hypothesis of homogeneity is rejected when $\chi_{\text{obs}}^2 > \chi_{1-\alpha; (I-1)(J-1)}^2$.

When row and column totals are not fixed, the numbers in the i, j cells can be used to estimate their corresponding population proportions without assuming the null hypothesis is true. With fixed row or column totals, this estimation cannot be accomplished. That is, $\hat{\pi}_{ij} = p_{ij} \neq \frac{n_{ij}}{n_{\bullet\bullet}}$ when H_0 is false.

Test for SCENARIO TWO:

Step 1: **Hypotheses** — $H_0 : \pi_{1j} = \pi_{2j}$ for all j versus $H_1 : \pi_{i,j} \neq \pi_{i+1,j}$ for some (i, j) .

That is, all the probabilities in the same column are equal to each other versus at least two of the probabilities in the same column are not equal to each other.

Step 2: **Test Statistic** — The test statistic is

$$\chi_{\text{obs}}^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi_{(I-1)(J-1)}^2 = \chi_{(2-1)(3-1)}^2 = \chi_2^2$$

under the null hypothesis. The χ_{obs}^2 value is 6.7584.

Step 3: **Rejection Region Calculations** — The rejection region is

$$\chi_{\text{obs}}^2 > \chi_{1-\alpha; (I-1)(J-1)}^2 = \chi_{0.95; 2}^2 = 5.991465.$$

Before the statistic $\chi_{\text{obs}}^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ can be computed, the expected counts for each of the ij cells must be determined. Recall that $O_{ij} = n_{ij}$ and $E_{ij} = \frac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}}$.

```
> DT <- c(67,76,57,48,73,79)
> DTT <- matrix(data=DT, nrow=2, byrow=TRUE)
> dimnames(DTT) <- list(Treatment=c("Drug", "Placebo"),
+ Category=c("Improve", "No Change", "Worse"))
> DTT
```

	Category		
Treatment	Improve	No Change	Worse
Drug	67	76	57
Placebo	48	73	79

```
> E <- outer(rowSums(DTT), colSums(DTT), "*")/sum(DTT)
> E
      Improve No Change Worse
Drug      57.5      74.5      68
Placebo   57.5      74.5      68
```

$$\chi_{\text{obs}}^2 = \frac{(67 - 57.5)^2}{57.5} + \frac{(76 - 74.5)^2}{74.5} + \dots + \frac{(79 - 68)^2}{68} = 6.76$$

The value of the test statistic is $\chi_{\text{obs}}^2 = 6.7584$. This can be done with code by entering

```
> chi.obs <- sum((DTT-E)^2/E )
> chi.obs
[1] 6.758357
```

$$6.76 = \chi_{\text{obs}}^2 \stackrel{?}{>} \chi_{.95,2}^2 = 5.99.$$

Step 4: **Statistical Conclusion** — The ϕ -value is 0.03407544.

```
> p.val <- 1-pchisq(chi.obs,2)
> p.val
[1] 0.03407544
```

- I. From the rejection region, since $\chi_{\text{obs}}^2 = 6.76 > \chi_{.95;2} = 5.99$, reject the null hypothesis of homogeneity.
- II. Since the ϕ -value = .034 is less than 0.05, reject the null hypothesis of homogeneity.

Reject H_0 .

Step 5: **English Conclusion** — There is sufficient evidence to suggest that not all of the probabilities for the $i = 2$ populations with respect to each of the J categories are equal.

Both R and S-PLUS have the function `chisq.test()`, which can be used to test the null hypothesis of homogeneity by computing the observed test statistic and its corresponding ϕ -value:

```
> chisq.test(DTT)
```

Pearson's Chi-squared test

```
data: DTT
X-squared = 6.7584, df = 2, p-value = 0.03408
```



10.9 Nonparametric Bootstrapping

The term “bootstrapping” is due to Efron (1979), and is an allusion to a German legend about a Baron Münchhausen, who was able to lift himself out of a swamp by pulling himself up by his own hair. In later versions he was using his own boot straps to pull himself out of the sea, which gave rise to the term bootstrapping. As improbable as it may seem, taking samples from the original data and using these **resamples** to calculate statistics can actually give more accurate answers than using the single original sample to calculate an estimate of a parameter. In fact, resampling methods require fewer assumptions than the traditional methods found in Chapters 7 and 8 and sometimes give more accurate answers. One of the more common methods of resampling is the bootstrap. The fundamental concept in bootstrapping is the building of a sampling distribution for a particular statistic by resampling from the data that are at hand. Although bootstrap methods are both parametric and nonparametric, attention in this section is focused exclusively on the nonparametric bootstrap. These methods offer the practitioner valuable tools for dealing with complex problems with computationally intensive techniques that rely on today’s computers, which are many times faster than those of a generation ago. Even though resampling procedures rely on the “new” power of the computer to perform simulations, they are based on the “old” statistical principles such as populations, parameters, samples, sampling variation, pivotal quantities, and confidence intervals.

For most students, the idea of a sampling distribution for a particular statistic is completely abstract; however, once work begins with the bootstrap distribution, the bootstrap analog to the sampling distribution, the concreteness of the bootstrap distribution promotes a conceptual understanding of the more abstract sampling distribution. In fact, bootstrap procedures promote statistical thinking by providing concrete analogies to theoretical concepts.

S-PLUS has very extensive built-in resampling capabilities. Two S packages for bootstrapping are **bootstrap** by Efron and Tibshirani (1993) (ported to R from S-PLUS by Friedrich Leisch) and **boot** by Angelo Canty (ported to R from S-PLUS by B. D. Ripley). Angelo Canty’s package provides functions and data sets from the book *Bootstrap Methods and Their Applications* by Davison and Hinkley (1997). The package **boot** is used for the remainder of this chapter. For R, **boot** can be obtained from CRAN; and for S-PLUS, from <http://statwww.epfl.ch/davison/BMA/library.html>.

10.9.1 Bootstrap Paradigm

Suppose a random sample $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$ is taken from an unknown probability distribution, F , and the values $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$ are observed. Using \mathbf{x} , the parameter $\theta = t(F)$ is to be estimated. The traditional approach of estimating θ is to make some assumptions about the population structure and to derive the sampling distribution of $\hat{\theta}$ based on these assumptions. This, of course, assumes the derivation of the sampling distribution of the statistic of interest has either been done or that the individual who needs to do the deriving has the mathematical acumen to do so. Often, the use of the bootstrap will be preferable to extensive mathematical calculations.

In the bootstrap paradigm, the original sample, \mathbf{x} , takes the place the population holds in the traditional approach. Subsequently, a random sample of size n is drawn from \mathbf{x} with replacement. The resampled values are called a bootstrap sample and are denoted \mathbf{x}^* .

These values are used to calculate an estimate of the statistic of interest, $s(\mathbf{x}) = \hat{\theta}$. This $s(\mathbf{x})$ is not necessarily the plug-in estimate of θ , $\hat{\theta} = t(\hat{F})$, where \hat{F} is the empirical probability distribution function. It is, however, the function applied to the bootstrap sample \mathbf{x}^* that creates a bootstrap estimate of θ denoted $\hat{\theta}^*$ or t^* . That is,

$$s(\mathbf{x}^*) = t^* = \hat{\theta}^*. \quad (10.33)$$

The star notation indicates that \mathbf{x}^* is not the original data set \mathbf{x} , but rather, it is a random sample of size n drawn with replacement from \mathbf{x} . That is, given $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$, one possible bootstrap sample \mathbf{x}^* might be $\{x_1^* = x_5, x_2^* = x_3, x_3^* = x_5, \dots, x_n^* = x_7\}$. Some values from the original sample \mathbf{x} may appear once, more than once, or not at all in the bootstrap sample \mathbf{x}^* .

The fundamental bootstrap assumption is that the sampling distribution of the statistic under the unknown probability distribution F may be approximated by the sampling distribution of $\hat{\theta}^*$ under the empirical probability distribution \hat{F} . That is,

$$\begin{aligned} \text{Var}_F(\hat{\theta}) &\approx \text{Var}_{\hat{F}}(\hat{\theta}^*) \\ G_F(a) &\approx G_{\hat{F}}(a) \\ G_F^{-1}(0.95) &\approx G_{\hat{F}}^{-1}(0.95) \end{aligned} \quad (10.34)$$

where G is the cumulative distribution function of the distribution of $\hat{\theta}$. Generally, the bootstrap estimate of the parameter of interest is not computed directly, but it is instead estimated from B bootstrap samples.

The process of creating a bootstrap sample \mathbf{x}^* and a bootstrap estimate $\hat{\theta}^*$ of the parameter of interest is repeated B times (typically 999 or more). The B bootstrap estimates of θ , the $\hat{\theta}^*$ s, are subsequently used to estimate specific properties of the bootstrap sampling distribution of $\hat{\theta}^*$. Note that B values of $\hat{\theta}^*$ are used to estimate specific properties of the bootstrap sampling distribution of $\hat{\theta}^*$. There are a total of $\binom{2n-1}{n}$ distinct bootstrap samples. Yet, a reasonable estimate of the standard error of $\hat{\theta}^*$, $\hat{\sigma}_{\hat{\theta}^*} \equiv \widehat{\text{SE}}_B$, can be achieved with only $B = 200$ bootstrap replications in most problems. For confidence intervals and quantile estimation, B generally should be at least 999.

The general procedure for estimating the standard error of $\hat{\theta}^*$ is

- (1) Generate B independent bootstrap samples $\{\mathbf{x}_1^*, \mathbf{x}_2^*, \dots, \mathbf{x}_B^*\}$, each consisting of n values drawn with replacement from \mathbf{x} .
- (2) Compute the statistic of interest for each bootstrap sample:

$$\hat{\theta}_b^* \equiv t_b^* = s(\mathbf{x}_b^*) \quad b = 1, 2, \dots, B$$

- (3) Estimate the standard error of $\hat{\theta}$ ($\text{SE}_F(\hat{\theta}) \equiv \sigma_{\hat{\theta}}$) by computing the sample standard deviation of the bootstrap replications of $\hat{\theta}_b^*$, $b = 1, 2, \dots, B$:

$$\sigma_{\hat{\theta}} \approx \widehat{\text{SE}}_B \equiv \hat{\sigma}_{\hat{\theta}^*} = \left[\sum_{b=1}^B \frac{(\hat{\theta}_b^* - \bar{\hat{\theta}^*})^2}{B-1} \right]^{\frac{1}{2}}, \quad \text{where} \quad \bar{\hat{\theta}^*} = \sum_{b=1}^B \frac{\hat{\theta}_b^*}{B}. \quad (10.35)$$

The bootstrap algorithm for estimating the standard error of a statistic $\hat{\theta} = s(\mathbf{x})$ is graphically depicted in Figure 10.16.

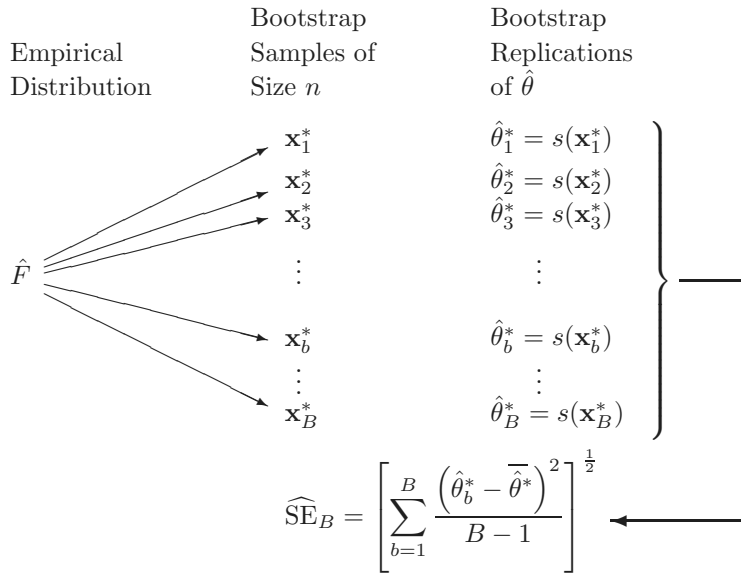


FIGURE 10.16: Graphical representation of the bootstrap based on Efron and Tibshirani (1993, Figure 6.1)

For most statistics, bootstrap distributions approximate the shape, spread, and bias of the actual sampling distribution; however, bootstrap distributions differ from the actual sampling distribution in the locations of their centers. The bootstrap distribution having a similar shape is clear. A similar spread means

$$Var[s(\mathbf{X})|F] \approx Var[s(\mathbf{X})|\hat{F}].$$

That is, the variance of the estimator $s(\mathbf{X})$ under the unknown distribution F is approximately the same as the variance of $s(\mathbf{X})$ under the bootstrap distribution obtained by replacing F with \hat{F} . The variance of $s(\mathbf{X})$ under \hat{F} is not computed, but rather estimated from the B bootstrap samples and is

$$Var[s(\mathbf{X})|\hat{F}] \approx \widehat{Var}_B[s(\mathbf{X})|\hat{F}] = \sum_{b=1}^B \frac{(\hat{\theta}_b^* - \bar{\theta}^*)^2}{B-1}.$$

The sampling distribution of a statistic $s(\mathbf{X})$ used to estimate the parameter $\theta = t(F)$ is centered at the parameter θ plus any bias, while the bootstrap distribution is centered at $\hat{\theta}$ plus any bias. Recall that the bias of a statistic $\hat{\theta}$ is $E(\hat{\theta}) - \theta$. Consequently, the bias of $s(\mathbf{X}) = \hat{\theta}$ is expressed as

$$Bias[s(\mathbf{X})|F] = E_F[s(\mathbf{X})] - t(F),$$

while the bias of the bootstrap distribution of $\hat{\theta}$ is

$$Bias[s(\mathbf{X})|\hat{F}] = E_{\hat{F}}[s(\mathbf{X}^*)] - t(\hat{F}).$$

Generally, the bootstrap bias of $s(\mathbf{X})$ is not computed directly but is instead estimated from B bootstrap samples. That is, $E_{\hat{F}}[s(\mathbf{X}^*)]$ is not computed directly but is estimated by

$$\hat{E}_{\hat{F}}[s(\mathbf{X}^*)] = \sum_{b=1}^B \frac{\hat{\theta}_b^*}{B} = \bar{\theta}^*.$$

The result is an estimated bootstrap bias of $s(\mathbf{X})$ based on B bootstrap samples denoted

$$\widehat{Bias}_B[s(\mathbf{X})] = \overline{\hat{\theta}^*} - \hat{\theta}. \quad (10.36)$$

10.9.2 Confidence Intervals

With estimates of the standard error (standard deviation) and bias of some statistic of interest, various types of confidence intervals for the parameter θ can be constructed. Although exact confidence intervals for specific problems can be computed, most confidence intervals are approximate. The most common confidence interval for a parameter θ when $\hat{\theta}$ follows either a normal or approximately normal distribution is

$$\hat{\theta} \pm z_{1-\alpha/2} \hat{\sigma}_{\hat{\theta}}. \quad (10.37)$$

The function `boot.ci()` in the package `boot` creates several types of confidence intervals, four of which are covered here.

The **normal** confidence interval in `boot.ci()` is a slight modification to (10.37) that incorporates both a bootstrap adjustment for bias and a bootstrap estimate of the standard error. The normal confidence interval is calculated as

$$CI_{1-\alpha}(\theta) = \left[\hat{\theta} - \widehat{Bias}_B(\hat{\theta}) - z_{1-\alpha/2} \cdot \widehat{SE}_B, \hat{\theta} - \widehat{Bias}_B(\hat{\theta}) + z_{1-\alpha/2} \cdot \widehat{SE}_B \right] \quad (10.38)$$

The **basic bootstrap** confidence interval is based on the idea that the quantity $\hat{\theta}^* - \hat{\theta}$ has roughly the same distribution as $\hat{\theta} - \theta$. Since (10.39) has $(\hat{\theta}^* - \hat{\theta}) \sim (\hat{\theta} - \theta)$, (10.40) follows. To get (10.41) from (10.40), subtract $\hat{\theta}$ inside the probability statement and divide by -1 :

$$\mathbb{P} \left[\hat{\theta}_{((B+1) \cdot \alpha/2)}^* - \hat{\theta} \leq \hat{\theta}^* - \hat{\theta} \leq \hat{\theta}_{((B+1) \cdot (1-\alpha/2))}^* - \hat{\theta} \right] \approx 1 - \alpha, \quad (10.39)$$

$$\mathbb{P} \left[\hat{\theta}_{((B+1) \cdot \alpha/2)}^* - \hat{\theta} \leq \hat{\theta} - \theta \leq \hat{\theta}_{((B+1) \cdot (1-\alpha/2))}^* - \hat{\theta} \right] \approx 1 - \alpha \quad (10.40)$$

$$\mathbb{P} \left[2\hat{\theta} - \hat{\theta}_{((B+1) \cdot (1-\alpha/2))}^* \leq \theta \leq 2\hat{\theta} - \hat{\theta}_{((B+1) \cdot \alpha/2)}^* \right] \approx 1 - \alpha \quad (10.41)$$

Equations (10.39) to (10.41) lead to the basic bootstrap confidence interval given in (10.42):

$$CI_{1-\alpha}(\theta) = \left[2\hat{\theta} - \hat{\theta}_{((B+1) \cdot (1-\alpha/2))}^*, 2\hat{\theta} - \hat{\theta}_{((B+1) \cdot \alpha/2)}^* \right] \quad (10.42)$$

The **percentile** confidence interval is based on the quantiles of the B bootstrap replications of $s(\mathbf{X})$. Specifically, the $(1 - \alpha)$ percentile confidence interval of θ uses the $\alpha/2$ and the $1 - \alpha/2$ quantiles of the $\hat{\theta}^*$ values to create a $(1 - \alpha) \cdot 100\%$ confidence interval for θ :

$$CI_{1-\alpha}(\theta) = \left[\hat{\theta}_{((B+1) \cdot \alpha/2)}^*, \hat{\theta}_{((B+1) \cdot (1-\alpha/2))}^* \right] \quad (10.43)$$

The notation $\hat{\theta}_{(\text{Integer})}^*$ is used to denote the $(\text{Integer})^{\text{th}}$ $\hat{\theta}^*$ of the B sorted $\hat{\theta}^*$ values. The values of B and α are generally chosen so that $(B + 1) \cdot \alpha/2$ is an integer. In cases where $(B + 1) \cdot \alpha/2$ is not an integer, interpolation can be used. (Note that different programs use different interpolation techniques.)

One may have noticed that the percentile confidence interval uses $\hat{\theta}_{((B+1)\cdot\alpha/2)}^*$ to construct the *lower* endpoint of the confidence interval while the basic bootstrap interval uses $\hat{\theta}_{((B+1)\cdot\alpha/2)}^*$ in the construction of the *upper* endpoint of its confidence interval. Is one of the methods backwards? If not, does one method work better than the other? In fact, neither method is backward and neither method is uniformly superior to the other. At this point, a reasonable question might be which confidence interval is recommended for general usage since the normal confidence interval is based on large sample properties and the percentile and basic bootstrap confidence interval formulas give different answers when the distribution of $\hat{\theta}^*$ is skewed. In fact, the answer is to use *none* of the confidence intervals discussed thus far. The bootstrap confidence interval procedure recommended for general usage is the BC_a method, which stands for bias-corrected and accelerated. The first three methods discussed (normal, percentile, and basic bootstrap) have first-order accuracy, while the BC_a method is second-order accurate. Accuracy in this context simply refers to the coverage errors. The bottom line is that there are theoretical reasons to prefer the BC_a confidence interval over the normal, percentile, and basic bootstrap confidence intervals.

To compute a BC_a interval for θ , $CI_{1-\alpha}(\theta) = [\hat{\theta}_{\text{lower}}^*, \hat{\theta}_{\text{upper}}^*]$, first compute the bias factor, z , where

$$z = \Phi^{-1} \left[\frac{\sum_{b=1}^B I\{\hat{\theta}_b^* < \hat{\theta}\}}{B} \right]. \tag{10.44}$$

Recall the definition of Φ^{-1} on page 153. Provided the estimated bootstrap distribution, $s(\mathbf{x}^*) = \hat{\theta}^*$, is symmetric with respect to $\hat{\theta}$, and if $\hat{\theta}$ is unbiased, then $\frac{\sum_{b=1}^B I\{\hat{\theta}_b^* < \hat{\theta}\}}{B}$ will be close to 0.5, and the bias correction factor z will be close to zero since $\Phi^{-1}(0.5) = 0$, with $S, \text{qnorm}(.5) = 0$. Next, compute the skewness correction factor:

$$a = \frac{\sum_{i=1}^n \left(\bar{\hat{\theta}}_{(-i)} - \hat{\theta}_{(-i)} \right)^3}{6 \left[\sum_{i=1}^n \left(\bar{\hat{\theta}}_{(-i)} - \hat{\theta}_{(-i)} \right)^2 \right]^{\frac{3}{2}}} \tag{10.45}$$

where $\hat{\theta}_{(-i)}$ is the value of $\hat{\theta} = s(\mathbf{X})$ when the i^{th} value is deleted from the sample of n values and $\bar{\hat{\theta}}_{(-i)} = \sum_{i=1}^n \frac{\hat{\theta}_{(-i)}}{n}$. Using z and a , compute

$$a_1 = \Phi \left[z + \frac{z + z_{\alpha/2}}{1 - a(z + z_{\alpha/2})} \right] \text{ and } a_2 = \Phi \left[z + \frac{z + z_{1-\alpha/2}}{1 - a(z + z_{1-\alpha/2})} \right] \tag{10.46}$$

Now, lower = $(B + 1) \cdot a_1$ and upper = $(B + 1) \cdot a_2$. When either lower or upper is not an integer, interpolation can be used to obtain the lower and upper endpoints of the BC_a confidence interval:

$$CI_{1-\alpha}(\theta) = [\hat{\theta}_{\text{lower}}^*, \hat{\theta}_{\text{upper}}^*]. \tag{10.47}$$

Let $k = \lfloor (B + 1) \cdot a_i \rfloor$ for $i = 1, 2$. Then the appropriate interpolation is

$$\hat{\theta}_{((B+1)\cdot a_i)}^* = \hat{\theta}_{(k)}^* + \frac{\Phi^{-1}(a_i) - \Phi^{-1}\left(\frac{k}{B+1}\right)}{\Phi^{-1}\left(\frac{k+1}{B+1}\right) - \Phi^{-1}\left(\frac{k}{B+1}\right)} \cdot \left(\hat{\theta}_{(k+1)}^* - \hat{\theta}_{(k)}^* \right) \text{ for } i = 1, 2. \tag{10.48}$$

Example 10.17 ▷ *Bootstrap CIs with M1 Motorspeedway Times* ◁ The times recorded are those for 41 successive vehicles traveling northwards along the M1 motorway in England when passing a fixed point near Junction 13 in Bedfordshire on Saturday, March 23, 1985. After subtracting the times, the following 40 interarrival times, reported to the nearest second, are stored in **SDS4** under the variable **Times**.

- (a) Determine the distribution of the interarrival times.
- (b) Calculate bootstrap confidence intervals for the mean and standard deviation of those times using the function `boot.ci()` from the `boot` package. Specifically, use the arguments `norm`, `basic`, `perc`, and `bca` with the function `boot.ci()` to create normal approximation, basic, percentile, and BC_a confidence intervals.
- (c) Verify the resulting confidence intervals for the mean using the appropriate equations.

Solution: The answers are as follows:

(a) It appears that the conditions for an approximate Poisson process are satisfied. The estimated parameter, λ , for the Poisson process, $\hat{\lambda}$, is 0.1282051 cars per second. Consequently, the waiting time until the next car follows an exponential distribution with mean $1/\lambda$. In this case, the estimated mean of the exponential distribution (waiting time) is 7.8 seconds/car. A density histogram of the interarrival times with a superimposed $Exp(\lambda = 0.1282051)$ suggests this distribution is reasonable. In addition, both the mean and the standard deviation of `Times` are roughly equal, as they should be with an exponential distribution.

```
> attach(SDS4)
> TT <- max(cumsum(Times))      # Time period (seconds)
> n <- length(lag(Times))      # Number of Cars
> Elamb <- n/TT                # Cars/Time period
> EMeanExp <- 1/Elamb          # Time period/Cars
> ans <- c(TT, n, Elamb, EMeanExp)
> names(ans) <- c("Total Time", "# of Cars", "Est. lambda", "Est. Mean")
> ans
  Total Time  # of Cars Est. lambda  Est. Mean
312.0000000  40.0000000  0.1282051  7.8000000
> hist(Times, prob=TRUE)
> curve(dexp(x, Elamb), 0, 35, add=TRUE) # Only R
> c(mean(Times), sd(Times))           # Distribution Check
[1] 7.800000 7.871402
```

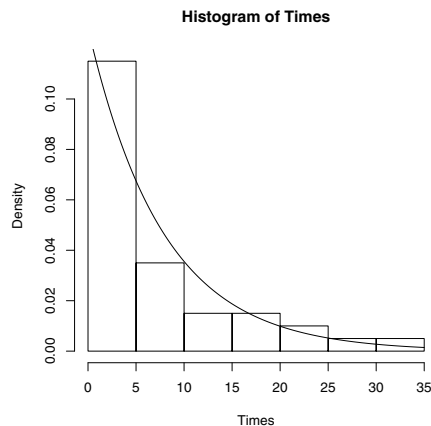


FIGURE 10.17: Histogram of interarrival times at the M1 motorspeedway

(b) The four bootstrap confidence intervals are now constructed for the mean using both the function `boot()` and the formulas. Use the function `Times.mean()` with the `boot` package to compute the mean:

```
> library(boot)
> Times.mean <- function(data, i)
+ {
+   d <- data[i]
+   M <- mean(d)
+   M
+ }
```

Set the number of bootstrap replications B to 9999 and generate the bootstrapped distribution of \bar{X} denoted by t^* when using the `boot` package. Store the results in `b.obj`. Note that `R` in `boot` is the number of bootstrap replications, which is denoted B in this text, so `R` is set equal to `B`. A random seed value of 10 (in `R`) is used so the reader can reproduce the results in the text:

```
> set.seed(10)
> B <- 9999
> b.obj <- boot(Times, Times.mean, R=B)
> b.obj
```

ORDINARY NONPARAMETRIC BOOTSTRAP

Call:

```
boot(data = Times, statistic = Times.mean, R = B)
```

Bootstrap Statistics :

	original	bias	std. error
t1*	7.8	-0.012494	1.215

Examine the graph (Figure 10.18) of t^* . Note that the histogram and normal quantile-quantile plot of t^* indicate the distribution is slightly skewed to the right. Hence, there are small differences between the confidence intervals created with the percentile method and the basic bootstrap.

```
> plot(b.obj) # Histogram and Q-Q plot of t* values
```

Next, use the function `boot.ci()` on the object `b.obj` to create the four types of bootstrapped confidence intervals:

```
> boot.ci(b.obj, type=c("norm","basic","perc","bca"))
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 9999 bootstrap replicates
```

CALL :

```
boot.ci(boot.out = b.obj, type = c("norm", "basic", "perc", "bca"))
```

Intervals :

Level	Normal	Basic
95%	(5.431, 10.194)	(5.275, 10.050)

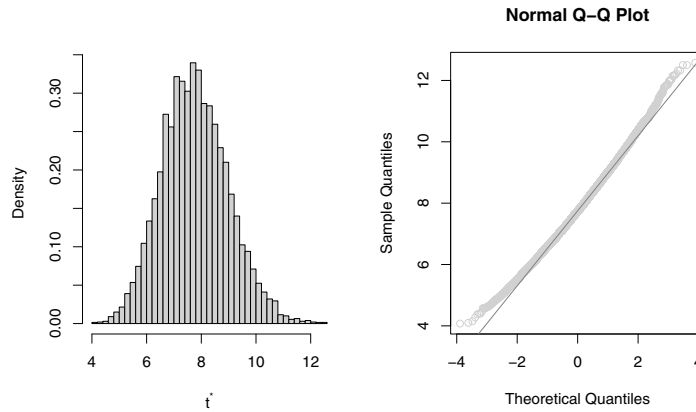


FIGURE 10.18: Histogram and quantile-quantile plot of the bootstrapped mean of interarrival times at the M1 motorspeedway

```

Level      Percentile      BCa
95% ( 5.550, 10.325 ) ( 5.800, 10.700 )
Calculations and Intervals on Original Scale

```

(c) To verify that these confidence intervals match the appropriate equations, start by setting $T_0 = t_{\text{obs}}$, $T_{\text{star}} = t^*$ values, and $\text{BIAS} = \text{estimated bootstrapped bias from (10.36)}$. A 95% confidence level matches the default confidence level for the function `boot.ci()`.

```

> T0 <- b.obj$t0
> Tstar <- b.obj$t
> BIAS <- mean(Tstar)-T0
> BIAS
[1] -0.012494
> conf.level <- .95
> alpha <- 1-conf.level

```

The normal confidence interval is calculated with (10.38). Note that the command `sd()` should be replaced with `stdev()` if using S-PLUS.

```

> c( (T0-BIAS)-qnorm(1-alpha/2)*sd(Tstar),
+   (T0-BIAS)+qnorm(1-alpha/2)*sd(Tstar) )
[1] 5.4312 10.1938

```

The basic confidence interval is calculated with (10.42).

```

> bt <- sort(Tstar)
> c(2*T0 - bt[(B+1)*(1-alpha/2)], 2*T0 - bt[(B+1)*(alpha/2)])
[1] 5.275 10.050

```

The percentile confidence interval is calculated using (10.43).

```

> c(bt[(B+1)*(alpha/2)], bt[(B+1)*(1-alpha/2)])
[1] 5.550 10.325

```

The BC_a confidence interval is figured with (10.47), including the bias factor z from (10.44), the a value in (10.45), and the a_1 and a_2 values in (10.46):

```
> z <- qnorm(sum(Tstar < T0)/B) # Using (10.44)
> z
[1] 0.045014
> n <- length(Times)
> u <- rep(0, n)
> for (i in 1:n)
+ {
+ u[i] <- mean(Times[-i])
+ }
> ubar <- mean(u)
> numa <- sum((ubar-u)^3)
> dena <- 6*sum((ubar-u)^2)^(3/2)
> a <- numa/dena
> a
[1] 0.043082
> a1 <- pnorm( z + (z-qnorm(1-alpha/2))/(1-a*(z-qnorm(1-alpha/2))) )
> a2 <- pnorm( z + (z+qnorm(1-alpha/2))/(1-a*(z+qnorm(1-alpha/2))) )
```

Since $(B+1) \cdot a_1$ and $(B+1) \cdot a_2$ are not integers, $\hat{\theta}_{\text{lower}}^*$ and $\hat{\theta}_{\text{upper}}^*$ are calculated with the interpolation (10.48):

```
> kl <- floor((B+1)*a1)
> ll <- bt[kl] + (qnorm(a1)-qnorm(kl/(B+1)))/(qnorm((kl+1)/(B+1))
+ -qnorm(kl/(B+1)))*(bt[kl+1]-bt[kl])
> ku <- floor((B+1)*a2)
> ul <- bt[ku] + (qnorm(a2)-qnorm(ku/(B+1)))/(qnorm((ku+1)/(B+1))
+ -qnorm(ku/(B+1)))*(bt[ku+1]-bt[ku])
> c(ll, ul)
[1] 5.8 10.7
```

Thus, all four intervals are verified to match the confidence intervals computed with `boot.ci()` by using their respective equations.

For the confidence intervals for the standard deviation, simply use `boot.ci()`. First write the function `Times.sd()` to compute standard deviation. Then, proceed with the calculation as before.

```
> Times.sd <- function(data, i)
+ {
+ d <- data[i]
+ SD <- sqrt(var(d))
+ SD
+ }
>
> set.seed(13)
> B <- 9999
> b.obj <- boot(Times, Times.sd, R=B)
```



```
> b.obj
```

```
ORDINARY NONPARAMETRIC BOOTSTRAP
```

```
Call:
```

```
boot(data = Times, statistic = Times.sd, R = B)
```

```
Bootstrap Statistics :
```

```
      original   bias  std. error
t1*    7.8714 -0.2209    1.2813
```

Construct a histogram and quantile-quantile plot of the t^* values. Note the slight negative skew in the distribution of t^* in Figure 10.19. This slight skewness will make the BC_a interval preferred over the other three types.

```
> plot(b.obj) # Histogram and Q-Q plot of t* values
```

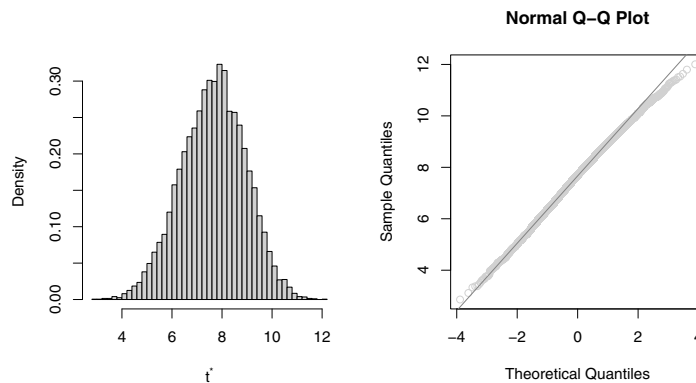


FIGURE 10.19: Histogram and quantile-quantile plot of the bootstrapped standard deviation of interarrival times at the M1 motorspeedway

Next, construct the four confidence intervals with `boot.ci()`:

```
> boot.ci(b.obj, type=c("norm","basic","perc","bca"))
```

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
```

```
Based on 9999 bootstrap replicates
```

```
CALL :
```

```
boot.ci(boot.out = b.obj, type = c("norm", "basic", "perc", "bca"))
```

```
Intervals :
```

```
Level      Normal              Basic
95%    ( 5.580, 10.603 )    ( 5.702, 10.659 )
```

Level	Percentile	BCa
95%	(5.084, 10.041)	(5.805, 10.891)

Calculations and Intervals on Original Scale ■

10.10 Permutation Tests

Permutation tests are computationally intensive techniques that actually predate computers. Until recently, permutation tests were more of a theoretical ideal than a useful technique. With the advent of high-powered computers, permutation tests have moved out of the abstract into the world of the practical. The permutation test is examined in only one context here—the two-sample problem. The fundamental idea behind the permutation test is that if there are no differences between two treatments, all data sets obtained by randomly assigning the data to the two treatments have an equal chance of being observed. Permutation tests are especially advantageous when working with small samples where verification of assumptions required for tests such as the pooled t -test are difficult.

To test a hypothesis with a permutation test:

- Step 1: Choose a test statistic $\hat{\theta}$ that measures the effect under study. Note that certain statistics will have more power to detect the effect of interest than others.
- Step 2: Create the sampling distribution that the test statistic in step 1 would have if the effect is not present in the population.
- Step 3: Find the “observed test statistic” in the sampling distribution from step 2. Observed values in the extremes of the sampling distribution suggest that the effect under study is “real.” In contrast, observed values in the main body of the sampling distribution imply that the effect is likely to have occurred by chance.
- Step 4: Calculate the φ -value based on the observed test statistic. This may be

$$\begin{aligned} \mathbb{P}(|\hat{\theta}| \geq |\hat{\theta}_{\text{obs}}|) \\ \mathbb{P}(\hat{\theta} \geq \hat{\theta}_{\text{obs}}) \\ \mathbb{P}(\hat{\theta} \leq \hat{\theta}_{\text{obs}}) \end{aligned}$$

for the inequality in H_A being \neq , $>$, or $<$, respectively.

The Two-Sample Problem

Suppose two independent random samples $\mathbf{z} = \{z_1, z_2, \dots, z_n\}$ and $\mathbf{y} = \{y_1, y_2, \dots, y_m\}$ are drawn from possibly different probability distributions F and G . The question of interest is whether $F = G$. Assuming $H_0 : F = G$ is true, it is possible to create the permutation sampling distribution for some statistic, $\hat{\theta}$, of interest. Let N equal the combined sample size $n + m$, let $\mathbf{v} = \{v_1, v_2, \dots, v_N\}$ be the combined and ordered vector of values, and let $\mathbf{g} = \{g_1, g_2, \dots, g_N\}$ be a vector indicating the group membership for \mathbf{v} . Since there are n z_i s and m y_j s in \mathbf{g} , there are $\binom{N}{n}$ possible ways of partitioning N elements into two subsets of sizes n and m . Consequently, under the null hypothesis that $F = G$, the vector \mathbf{g} has probability $1/\binom{N}{n}$ of equalling any one of its possible values. That is, all combinations of z_i s and y_j s are equally likely if $F = G$.

Suppose $H_A : F > G$ and $\hat{\theta} = \bar{z} - \bar{y}$. Then the exact φ -value is found by finding $\#\{\hat{\theta} \geq \hat{\theta}_{\text{obs}}\} / \binom{N}{n}$. For all but relatively trivial sized samples n and m , the number $\binom{N}{n}$ will be huge, making the enumeration of all possible samples of the statistic of interest a monumental task. Consider that, if $n = 10$ and $m = 10$, complete enumeration would require listing $\binom{20}{10} = 184,756$ possible outcomes.

Consequently, an approximation to the exact φ -value is often obtained by resampling without replacement the original data some “large” number of times, B , which is usually at least 999, and approximating the φ -value with

$$\varphi\text{-value} \approx \frac{1 + \#\{\hat{\theta}_b^* \geq \hat{\theta}_{\text{obs}}\}}{(B + 1)} = \frac{1 + \sum_{i=1}^B \mathbf{I}\{\hat{\theta}_b^* \geq \hat{\theta}_{\text{obs}}\}}{(B + 1)} \quad (10.49)$$

The resampling is done without replacement to approximate the φ -value from a permutation test. When the sampling is done with replacement, a bootstrap test is performed. Bootstrap tests are somewhat more general than permutation tests since they apply to a wider class of problems; however, they do not return “exact” φ -values.

Example 10.18 ▷ *Permutation Test* ◁ The data set used in this problem (`Ratbp`) is originally from Ott and Mendenhall (1985, problem 8.17). Researchers wanted to know whether a drug was able to reduce the blood pressure of rats. Twelve rats were chosen and the drug was administered to six rats, the treatment group, chosen at random. The other six rats, the control group, received a placebo. The drops in blood pressure (mmHg) for the treatment group (with probability distribution F) and the control group (with probability distribution G) are stored in the variables `Treat(z)` and `Cont(y)`, respectively. Note that positive numbers indicate blood pressure decreased while negative numbers indicate that it rose. Under the null hypothesis, $H_0 : F = G$, the data come from a single population. The question of interest is, “How likely are differences as extreme as those observed between the treatment and control groups to be seen if the null hypothesis is correct?” Use $\hat{\theta} = \bar{z} - \bar{y}$ as the statistic of interest and compute:

- The exact permutation φ -value,
- An estimated permutation φ -value based on 499 permutation replications, and
- An estimated bootstrap φ -value based on 499 bootstrap replications.

Solution: The test statistic of interest (step 1) has been specified to be $\hat{\theta} = \bar{z} - \bar{y}$. Finding the φ -values requires the creation of sampling distributions with different methods. First, load the `boot` package and attach `Ratbp`. Combine the treatment and control data in a single variable called `Blood`.

```
> library(boot)
> B <- 999
> attach(Ratbp)
> Ratbp
  Treat Cont
1  69.0   9.0
2  24.0  12.0
3  63.0  36.0
4  87.5  77.5
5  77.5  -7.5
6  40.0  32.5
> Blood <- c(Treat, Cont)
```

(a) **Exact Permutation Method** The variable `pdT6` contains all of the possible combinations, $\binom{12}{6} = 924$ of them, of the indices of `Blood`. `Comb` is a 924×12 matrix with each row containing the data in `Blood`. `Theta` will contain all possible values of the statistic $\hat{\theta} = \bar{z} - \bar{y}$. Note that an array of 924 zeros is `Theta`'s initial value before the `for` loop. The assignment following `Theta[i]`, `mean(Comb[i, pdT6[i,]]) - mean(Comb[i, -pdT6[i,]])`, takes each row of the `Comb` matrix and finds the value of $\hat{\theta}$ for the appropriate permutation using the indices in `pdT6`. For example, when $i = 751$, `pdT6[751,]` is (1, 2, 4, 8, 11, 12), and for

```
Comb[751,]=Blood= (69.0, 24.0, 63.0, 87.5, 77.5, 40.0,
                  9.0, 12.0, 36.0, 77.5, -7.5, 32.5),
```

`Comb[i, pdT6[i,]]` is `Comb[751, c(1,2,4,8,11,12)]`, which extracts

```
(69, 24, 87.5, 12, -7.5, 32.5)
```

and `Comb[i, -pdT6[i,]]` is `Comb[751, c(3,5,6,7,9,10)]`, which extracts

```
(63, 77.5, 40, 9, 36, 77.5).
```

This puts the mean of (69, 24, 87.5, 12, -7.5, 32.5) = 36.25 minus the mean of (63, 77.5, 40, 9, 36, 77.5) = 50.5, which is -14.25 in `Theta[751]`. The `for` loop calculates all $\binom{12}{6} = 924$ values of $\hat{\theta}$ in a similar manner. `Theta.obs` is the actual observed value of $\hat{\theta} = 60.167 - 26.583 = 33.583$, the mean blood pressure drop of the treatment group minus the mean blood pressure drop of the control group. The exact ϕ -value is the number of values in `Theta` greater than or equal to 33.583 divided by 924. This value is 0.031385, the exact permutation ϕ -value.

```
> pdT6 <- SRS(1:12,6) # OR t(Combinations(12,6))
> Comb <- matrix(rep(c(Treat, Cont),924), ncol=12, byrow=TRUE)
> Theta <- array(0,924)
> for(i in 1:924)
+ {
+ Theta[i] <- mean(Comb[i, pdT6[i,]]) - mean(Comb[i, -pdT6[i,]])
+ }
> Theta.obs <- mean(Treat) - mean(Cont)
> pval <- sum(Theta >= Theta.obs)/choose(12,6)
> pval
[1] 0.031385
```

Figure 10.20 shows a histogram of the sampling distribution of $\hat{\theta} = \bar{z} - \bar{y}$ with a vertical line marking $\hat{\theta}_{\text{obs}}$. The R function `oneway_test()` from the `coin` package can also be used to get an exact answer:

```
> library(coin)
> GR <- as.factor(c(rep("Treat",6), rep("Cont",6)))
> oneway_test(Blood~GR, distribution="exact", alternative="less")
```

Exact 2-Sample Permutation Test

```
data: Blood by groups Cont, Treat
Z = -1.871, p-value = 0.03139
alternative hypothesis: true mu is less than 0
```

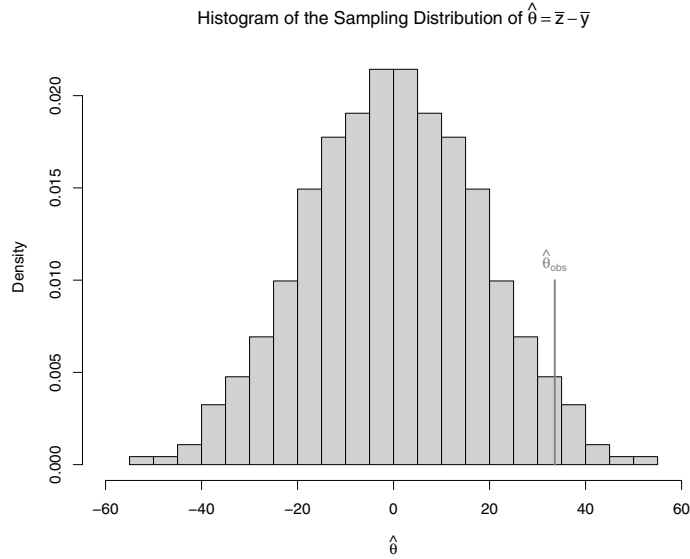


FIGURE 10.20: Histogram of the sampling distribution of $\hat{\theta} = \bar{z} - \bar{y}$

(b) **Estimated Permutation Method** based on $B=499$ resamples of `Blood` without replacement. After assigning 499 to `B`, the function `blood.boot()` is created to compute the mean difference between the first six values and the last six values in a vector of length twelve. Note that the object `data` is what will be resampled. To resample without replacement, the argument `sim="permutation"` is used with `boot()`. The φ -value is computed according to (10.49) and the estimated permutation φ -value based on $B=499$ permutation replications is 0.03. A histogram and a quantile-quantile plot of $\hat{\theta}^* = \bar{z}^* - \bar{y}^*$ when sampling without replacement are shown in Figure 10.21 on the facing page.

```
> blood.fun <- function(data, i)
+ {
+   d <- data[i]
+   MD <- mean(d[1:6]) - mean(d[7:12])
+   MD
+ }
> set.seed(13)
> B <- 499
> boot.blood <- boot(Blood, blood.fun, R=B, sim="permutation")
> plot(boot.blood)
> pval.boot <- (sum(boot.blood$t >= boot.blood$t0)+1)/(B+1)
> pval.boot
[1] 0.03
```

(c) **Estimated Bootstrap Method** based on $B=499$ resamples of `Blood` with replacement. After assigning 499 to `B`, the function `blood.boot()` is used with `boot()` to create an estimated bootstrap distribution based on $B=499$ resamples. Note that the argument `sim="ordinary"` is used with `boot()` to sample with replacement. The φ -value is computed according to (10.49) and the estimated bootstrap φ -value based on $B=499$ replications is 0.038. A histogram and a quantile-quantile plot of $\hat{\theta}^* = \bar{z}^* - \bar{y}^*$ when sampling with replacement are shown in Figure 10.22 on the next page.

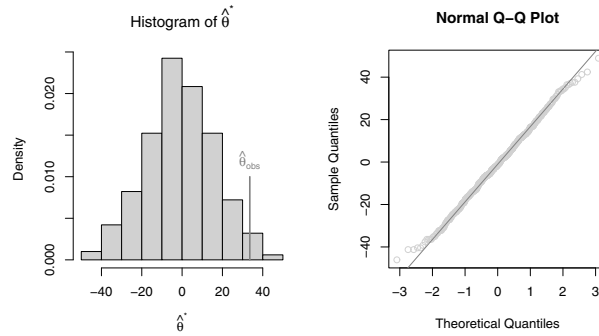


FIGURE 10.21: Histogram and quantile-quantile plot of $\hat{\theta}^* = \bar{z}^* - \bar{y}^*$ when sampling without replacement (permutation)

```
> set.seed(13)
> B <- 499
> boot.blood.b <- boot(Blood, blood.fun, R=B, sim="ordinary")
> plot(boot.blood.b)
> pval.boot.b <- (sum(boot.blood.b$t >= boot.blood.b$t0)+1)/(B+1)
> pval.boot.b
[1] 0.038
```

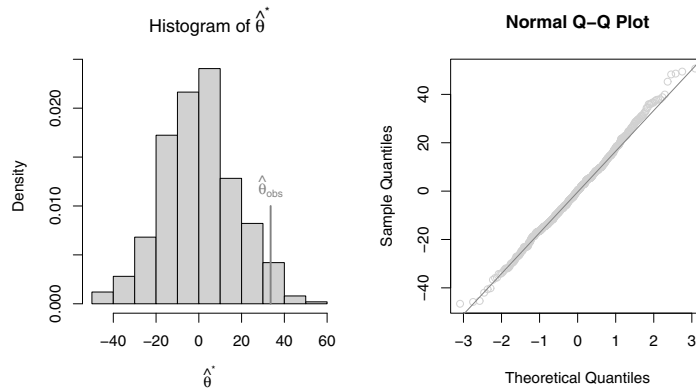


FIGURE 10.22: Histogram and quantile-quantile plot of $\hat{\theta}^* = \bar{z}^* - \bar{y}^*$ when sampling with replacement (bootstrap) ■

10.11 Problems

1. Provide a brief explanation of the pros and cons of using a nonparametric test.
2. What are the assumptions made with respect to the distribution from which the data come when
 - (a) using a sign test?
 - (b) using the Wilcoxon signed-rank test?
3. When testing the median difference (ψ_D) of two dependent samples, does the sign test or the Wilcoxon signed-rank test have more power? For the recommended test, what assumption(s) must be made?
4. Explain when it might be appropriate to use
 - (a) a Kruskal-Wallis test.
 - (b) a Friedman Test.
5. Explain the concept “goodness-of-fit” as used for tests in this chapter.
6. The service department of an automobile dealer is being evaluated on the quality of their service. The parent company has randomly chosen 11 clients who have had their vehicles serviced in the last six months. These clients have been asked to evaluate their satisfaction level with the service they received. The following satisfaction scores were obtained:

Scores: 6 10 8 3 6 2 8 9 10 10 2

Service departments where the company has evidence that the median rating is more than 7 will receive a bonus. Perform the appropriate hypothesis test to determine if this service department should be awarded a bonus.

7. A Mendebaldea real estate agent claims Mendebaldea, Spain, has larger apartments than those in San Jorge, Spain. A San Jorge real estate agent disputes this claim. To resolve the issue, two random samples of the total area of several apartments (given in m²) are taken from each community in 2002 and stored in the data frame `AptSize`.

Mendebaldea	90	92	90	83	85	105	136
San Jorge	75	75	53	78	52	90	78

- (a) Is there evidence to support the Mendebaldea agent’s claim?
 - (i) Use an exact procedure.
 - (ii) Use an approximate procedure.
 - (b) Find a confidence interval for the median difference (Mendebaldea–San Jorge) with a confidence level of at least 0.90.
8. To study the retained carbon of trees, a sample of 41 plots has been drawn in different mountainous regions of Navarra (Spain). In these plots, the carbon retained by leaves has been measured in kg/ha, depending on the forest classification: Areas with 90%

or more beech trees (*Fagus Sylvatica*) are labeled monospecific, while areas with many species of trees are labeled multispecific. The data are stored in the data frame `fagus`. Is there evidence that leaves from different forest classifications retain the same amount of carbon?

SOURCE: Data come from *Gobierno de Navarra* and *Gestión Ambiental de Viveros y Repoblaciones de Navarra*, 2006. The data were obtained within the European Project FORSEE.

9. The R data set `USJudgeRatings` provides 43 lawyers' ratings of state judges serving in the U.S. Superior Court. Use `help(USJudgeRatings)` to obtain a detailed view of the file. Suppose the variables `integrity` (`INTG`) and `demeanor` (`DMNR`) are chosen.
 - (a) Test whether lawyers are more likely to give a judge high integrity ratings rather than high demeanor ratings.
 - (b) Find a confidence interval for the median difference (`integrity`–`demeanor`) with a confidence level of at least 0.90.
10. A company manager is studying the possibility of giving 20 minutes of rest to her employees in a resting room. To check the viability of this proposal, she analyzed 12 days of productivity where employees took 20 minutes of rest and 12 days where they did not. The employee productivity scores are given in the following table where higher scores represent greater productivity.

With Rest	9	8	8	7	6	7	8	9	7	7	7	6
Without Rest	7	9	5	6	7	3	9	9	4	5	6	4

Is there evidence to suggest that taking a rest produces an increase in median employee productivity?

11. A Japanese company and an American company each claims that it has developed new technology to increase network transmission speeds. The marketing managers of both companies simultaneously announce that they can transmit 1 terabyte per second. To substantiate their claims, each company submits trial data (in seconds) to transmit one terabyte with the new technologies:

Japanese company	0.98	0.85	0.9	1	
American company	0.95	0.94	0.8	1	0.99

Is there evidence to suggest the transmission speed using the technology developed by the American company is superior to the transmission speed using the technology developed by the Japanese company? Compute the ϕ -value to answer the question with the following techniques:

- (a) Enumerate all possible combinations with the function `SRS()` to find the ϕ -value for a permutation test.
- (b) Use the function `oneway_test()` from the `coin` package to calculate an appropriate ϕ -value. Does this ϕ -value match the one in part (a)?
- (c) Obtain an estimated permutation ϕ -value using the `boot()` function from the `boot` package.
- (d) What conclusion do the ϕ -values support?

12. The R data frame `sleep` shows the increase or decrease in hours of sleep for two groups of patients when compared with a control group. Both groups were provided with a different soporific drug. Is there evidence to suggest that one drug is superior (induces more sleep) than the other drug?
13. In 1876, Charles Darwin had his book, *The Effect of Cross and Self-Fertilization in the Vegetable Kingdom*, published. Darwin planted two seeds, one obtained by cross-fertilization and the other by auto-fertilization in two opposite but separate locations of a pot. Self-fertilization, also called autogamy or selfing, is the fertilization of a plant with its own pollen. Cross-fertilization, or allogamy, is the fertilization with pollen of another plant, usually of the same species. Darwin recorded the plants' heights in inches. The data frame `Fertilize` from the `PASWR` package contains the data from this experiment.
- Are the samples independent or paired?
 - Should normality be assumed?
 - Do the more appropriate test to decide if significance differences exist between the median heights of the plants with regard to fertilization methods.
14. Salaries for graduates of three engineering universities ten years after graduation are provided in the data frame `Engineer` of the `PASWR` package. Seventeen graduates were randomly selected from each university, and their salaries in thousands of dollars were recorded. Is there any evidence to suggest graduates earn different salaries based on the university from which they graduated?
15. An engineering team is studying four different circuits that regulate the light intensity of a conference room. An accelerated life test was used to estimate the lifetime of each circuit. The results (lifetimes in thousands of hours) are stored in the data frame `CircuitDesigns` and are

Design 1	3.07	1.20	0.95	1.38	5.48	1.19	
Design 2	0.33	0.60	0.39	2.05	0.25	1.71	1.93
Design 3	1.75	2.41	2.02	2.24	1.69	1.24	
Design 4	2.03	3.50	1.95	3.09	2.90	2.37	3.20

- Do an exploratory analysis of the data and decide if normality can be assumed.
 - Choose an appropriate test to decide if there exist significant differences among the circuit designs. Use $\alpha = 0.05$.
16. The R data frame `airquality` shows daily air quality measurements in New York City, NY, from May to September 1973.
- Attach the `airquality` data, and read the definition of the variables with the command `help(airquality)`. What does the symbol `NA` mean?
 - Create a boxplot and a density plot of `Month` versus `ozone` as in Example 10.10 on page 437. Do the graphs for each month exhibit similar shapes?
 - Is it reasonable to assume that each month has a normally distributed ozone level?
 - Is there evidence that differences exist in the ozone levels during the year?
17. The R data frame `warpbreaks` gives the number of warp breaks per loom, where a loom corresponds to a fixed length of yarn.

- (a) Attach the `warpbreaks` data and use the function `xtabs()` to create a contingency table containing the number of warp breaks classified by `wool` and `tension`.
- (b) Is there an association between wool type and tension level?
18. The music industry wants to know if the musical style on a CD influences how many illegal copies of it are sold. To achieve this purpose, the company chooses six cities randomly and writes down the number of illegal CDs available on the street categorized by music type: classic music, flamenco, heavy-metal, and pop-rock. The data are shown in the following table.

City	Musical Style			
	Classical	Flamenco	Heavy-Metal	Pop-Rock
City 1	4	1	6	9
City 2	3	4	5	10
City 3	2	1	8	14
City 4	5	3	2	7
City 5	2	3	6	14
City 6	9	1	2	6

- (a) Use lattice/Trellis functions to create a box and whiskers plot and a density plot of the number of illegal CDs available for each music style.
- (b) Are the distribution shapes similar?
- (c) Are there significant differences in the numbers of CDs available according to musical style?
19. A regulatory commission is investigating whether an association exists between the number of branches of certain banks and the regions where they are located. The branches belong to the following banks: Bilbao-Vizcaya (BBVA), Caja Madrid (CM), La Caixa (LC), and Banco Santander (BS). The regions are Navarra, Alava, Guipuzcoa, and Vizcaya. The number of branches classified by banks and regions follow.

Province	Bank			
	BBVA	CM	LC	BS
Navarra	47	8	54	43
Álava	31	5	21	17
Guipuzcoa	64	4	43	43
Vizcaya	134	11	104	66

Is there evidence that an association exists between region and number of branches?

20. The data frame `Depend` from the PASWR package shows the number of dependent children (`number`) for 50 families (`count`). Use a goodness-of-fit test to see if a Poisson distribution with $\lambda = 2$ can reasonably be used to model the number of dependent children.
21. Is it reasonable to assume that the `time` variable from the data frame `Phone` in the PASWR package follows an exponential distribution?
22. The data frame `TestScores` in the PASWR package gives the test grades of 29 students taking a basic statistics course.
- (a) Use the function `EDA()` on the data. Can normality be assumed?

- (b) Use a Kolmogorov-Smirnov test to assess normality.
23. A government grant is funding a study to calculate how long it takes for the average consumer to establish an Internet connection. A random sample of 20 Internet users' connection times is collected. The connections times in seconds are 0.03, 0.48, 0.49, 0.52, 0.66, 0.69, 0.70, 0.76, 0.82, 1.20, 1.22, 1.39, 1.62, 1.85, 1.97, 2.25, 2.84, 3.44, 3.48, and 4.02.
- (a) Use a Kolmogorov-Smirnov test to see if it is reasonable to assume that Internet connection time follows an exponential distribution with a mean of 1.5 seconds.
- (b) Use a chi-square test to see if it is reasonable to assume that Internet connection time follows an exponential distribution with a mean of 1.5 seconds.
- (c) Is there evidence to suggest that the median connection time is greater than 1 second?
24. Perform a simulation study to determine the power of both Kolmogorov-Smirnov's and Shapiro-Wilk's normality tests.
- (a) Set the seed equal to 897, and simulate $m = 10,000$ samples of size $n = 10, 20, 30,$ and 40 from a χ_1^2 distribution. Compute the simulated power for both tests using $\alpha = 0.05$.
- (b) Set the seed equal to 897, and simulate $m = 10,000$ samples of size $n = 10, 20, 30,$ and 40 from a $Unif(0, 1)$ distribution. Compute the simulated power for both tests using $\alpha = 0.05$.
- (c) Set the seed equal to 897, and simulate $m = 10,000$ samples of size $n = 10, 20, 30,$ and 40 from a $\beta(8, 3)$ distribution. Compute the simulated power for both tests using $\alpha = 0.05$.
- (d) Set the seed equal to 897, and simulate $m = 10,000$ samples of size $n = 10, 20, 30,$ and 40 from a $N(0, 1)$ distribution. Compute the simulated power for both tests using $\alpha = 0.05$.
- (e) Generalize your findings from (a) through (d).
25. The R data frame `HairEyeColor` contains classifications of 592 students by gender, hair color, and eye color.
- (a) Is hair color independent of eye color for men?
- (b) Is hair color independent of eye color for women?
26. The sinking of the *Titanic* occurred on the 15th of April in 1912. The data frame `titanic3` contains information regarding class, gender, and survival as well as several other variables.
- (a) Create contingency tables of
- passenger class versus survival,
 - male passengers' class versus survival, and
 - female passengers' class versus survival.
- (b) Is there an association between class and survival for all passengers, men, and/or women?

27. Mental inpatients in the Virgen del Camino Hospital (Pamplona, Spain) are interviewed by expert psychiatrists to diagnose their illnesses. An important aspect in diagnosis is determining the severity of any delusions a patient might suffer. A new questioning technique has been developed to detect the presence of delusions. The technique assigns a score from 0 to 5, where 5 indicates the presence of strong delusions and a 0 indicates no delusions. The psychiatrists wish to know if the new technique actually results in high scores for patients who have previously been diagnosed as suffering from severe delusions. The scores that follow were obtained from randomly selected patients who were known to suffer from delusions and those who were known not to suffer with delusions:

	Score						
Delusions Present	5	5	4	5	4	5	5
Delusions Absent	1	0	5	0	4	4	0

Do the data provide evidence that the new test yields higher scores for those patients who are known to suffer from delusions than for those who do not suffer from delusions?

28. It is believed by conservative psychiatrists that the use of illegal drugs can produce persistent hallucinations, even after drug use stops. Some more liberal psychiatrists dispute this assertion. The following data rate the severity of hallucinations suffered by randomly selected mental inpatients from the Virgen del Camino Hospital (Pamplona, Spain), where a 5 indicates severe hallucinations and a 0 indicates no hallucinations. The patients are divided by whether or not they consumed illegal drugs before being admitted to the hospital.

	Score							
Illegal Drugs Not Consumed	2	0	0	5	5	2	4	0
Illegal Drugs Consumed	0	4	5	5	4	5	5	2

Is there evidence that the severity of hallucinations in patients who have consumed illegal drugs is greater than the severity of hallucinations in patients who have not consumed illegal drugs?

29. Generate 10 values from a $N(0, 1)$ distribution with the seed set at 10. Calculate a bootstrap estimation of the standard error of \bar{X} using $B = 200$ replications. Repeat the experiment generating a sample of size 100 from the standard normal. What conclusions can be drawn?

Chapter 11

Experimental Design

11.1 Introduction

This chapter deals with designed experiments where the experimenter follows a specific protocol established before the experiment starts. This protocol should dictate how randomization is performed and how measurements are taken. As a consequence of adhering to an established protocol, designed experiments allow the user to make strong inferences about the nature of observed differences.

Experiments are generally conducted to compare groups in terms of some response of interest. The methods considered in this chapter assume the response variable is continuous. The factors, independent variables whose levels are set by the experimenter, are categories or continuous variables that have been categorized into a fixed number of discrete levels. The treatments of an experiment are applied to **experimental units**, and measurements on the response variable are taken where the objective of the experiment is to compare the observed responses. When the combinations of the levels of two or more factors form the treatments of interest, the experiment is known as a **factorial design**.

For example, an agricultural researcher may be interested in determining which of three different fertilizers produces the greatest soybean yield. In this example, the three fertilizers correspond to three treatments the experimenter wants to compare, and the three fertilizers collectively constitute a factor. In the event a second factor, such as two different methods of watering the soybeans, is of interest, the experiment will consist of $3 \times 2 = 6$ different treatment combinations and is called a factorial design.

Suppose an agronomist is interested in determining which of three types of wheat (*Triticum aestivum*, *Triticum durum*, or *Triticum spelta*) produces the greatest yield for a particular geographical location. Available to the agronomist are six plots of equal size, all in the same geographical area. In this setting, the factor of interest is wheat, and the treatments are the three types of wheat. When the plots are homogeneous in their physical characteristics, distinguishing differences in treatments becomes easier if differences exist. Since there are likely to be some differences in the plots, the researcher will want to assign the wheat types to the six plots randomly in order to minimize any possible bias due to plots. By randomizing the assignment of treatments, the possibility of confounding differences due to types with differences due to plots is minimized. When the assignment of treatments is done in a completely random fashion, the design is known as a completely randomized design (CRD). When experimental units are similar (homogeneous) with respect to some characteristic, they can be grouped together into **blocks**. In the wheat study, some of the plots may be exposed to more sun than other plots, or some plots may receive more water than other plots. When the experimental units are more homogeneous within a block than they are between blocks, treatments are assigned to experimental units within each block to reduce variability. Such a design is known as a randomized complete block design (RCBD). See Figures 11.1 and 11.2 on the next page for possible assignments of treatments for a CRD and RCBD, respectively.

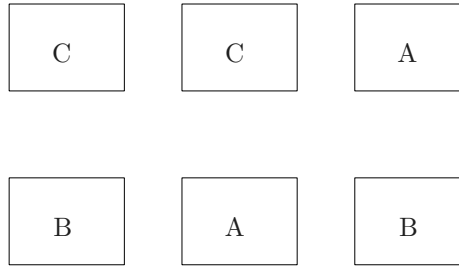


FIGURE 11.1: Representation of a completely randomized design where treatments A, B, and C are assigned at random to six experimental units.

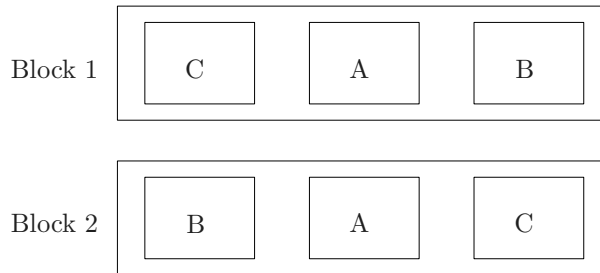


FIGURE 11.2: Representation of a randomized complete block design where treatments A, B, and C are assigned at random to three experimental units in each block.

Before proceeding further, some of the more important experimental design concepts are defined:

- **Treatments** are levels of a factor or combinations of factor levels the experimenter wants to compare.
- **Experimental units** are anything to which treatments are applied, for example, animals, plots, plants, or people.
- **Responses** are outcomes observed after the application of a treatment to an experimental unit.
- **Experimental error** is random variation present in the experiment not under the control of the experimenter. Experimental error may be due to many things, including but not limited to: measurement error, different responses from measuring the same quantity in different trials, and different responses from experimental units given the same treatment.
- **Treatment structure** specifies the set of factors the experimenter has selected to study or compare.
- **Design structure** defines how experimental units are assigned to treatment groups.
- **Randomization** is the use of some well-defined probabilistic mechanism to assign treatments to experimental units. Randomization reduces the possibility of bias and confounding. Randomization should also be used, if possible, with any variable not under the direct control of the experimenter that may influence the measured response.

- **Replication** is the independent assignment of several experimental units to each treatment (factor combination) resulting in independent observations. Replication shows the results are reproducible and allows the experimenter to estimate the experimental error. When the number of experimental units is the same for all treatments, the design is referred to as a balanced design. Unbalanced designs do not have an equal number of experimental units for all treatments.

Understanding both the treatment structure and the design structure is essential for conducting proper data analysis. Different statistical models will be introduced throughout the chapter, and analysis of variance (ANOVA) will be used. ANOVA measures the differences in means due to the factors that are fixed effects. When the effects are random, variance components are used to determine the variability due to the factors. All of the fixed effects models assume:

1. The measured responses are independent of one another.
2. The model errors are independent of one another and follow a normal distribution.
3. The variance is homogeneous across treatments.

When using statistical models, it is important to keep in mind that a model is simply a mathematical expression of how the researcher believes the response is explained using the independent variables of the experiment (predictors). Models are expressed in **S** with the syntax `response ~ predictors`, where `~` means that the `response` is modeled by the `predictors`. There may be several plausible models for a particular experiment. Finding an adequate model is an iterative process that starts by:

1. Identifying an appropriate model based on the treatment and design structure of the experiment.
2. Validating the model's assumptions using diagnostic plots.
3. Selecting a different model or transforming the response variable when the model's assumptions are not satisfied until a plausible model is found.

Once a model has been validated, formal inference to test for no treatment effects (equality of treatment means) and estimation of the model's parameters can be undertaken. In the event formal inference suggests differences in treatments, multiple comparisons are used to determine which treatments are significantly different from one another.

Motivational Example: Tires A tire manufacturer is interested in investigating the handling properties for different tread patterns. The data frame `Tire` has the stopping distances measured to the nearest foot for a standard sized car to come to a complete stop from a speed of 60 miles per hour. There are six measurements of the stopping distance for four different tread patterns labeled A, B, C, and D. The same driver and car were used for all 24 measurements. While the numbers in `Tire` do not reveal the randomization scheme used for the experiment, the order of treatments was assigned at random.

One way to ensure treatments are randomly assigned to the 24 runs is to use a random number generator. This can be accomplished with **S** by typing

```
> population <- rep(LETTERS[1:4], 6)
> Treatment <- sample(population)
> datf <- data.frame(Run=1:24, Treatment)
> datf[1:5,]
```


Run	Treatment
1	A
2	D
3	B
4	C
5	B

In particular, this randomization would assign tire tread A to the first run, D to the second run, and so on.

It is always a good idea to examine experimental data graphically before initiating any formal inferential procedure. Side-by-side boxplots are often a good starting point when comparing several treatments. When the number of observations in each treatment group is relatively small, dotplots will often prove more helpful than boxplots. The function `oneway.plots()` from the `PASWR` package is used to create Figure 11.3.

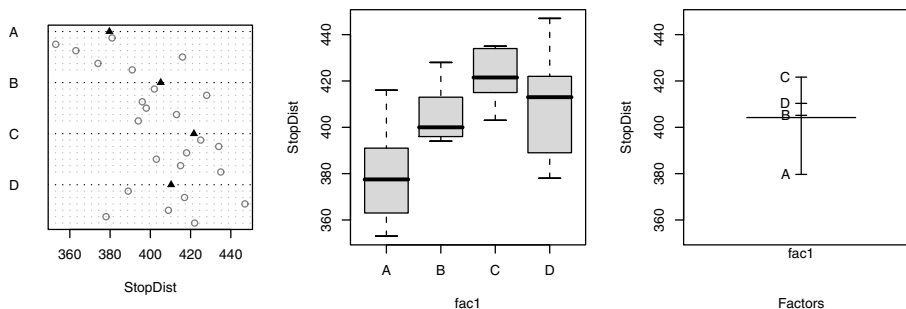


FIGURE 11.3: Output from the function `oneway.plots(StopDist, tire)` including dotplot, boxplots, and design plot (means) using the data frame `Tire`

From the boxplots and dotplots shown in Figure 11.3, it appears that there are differences in stopping distances based on different tire treads. At this point, it would be nice to formalize the last sentence with an inferential procedure. It is tempting to many to perform pairwise t -tests on all six $\binom{4}{2} = 6$ of the pairwise differences; however, this should not be done! If the probability of correctly accepting the null hypothesis is $1 - \alpha = 0.90$, then the probability of correctly accepting the null hypothesis for all six pairwise tests assuming independence among tests would be $(.90)^6 = 0.7350919$. The type I error rate is not 10% but 26.5% in this case. Of course, the more treatments that are compared, the more likely one is to make a type I error. What would the type I error rate be if the individual error rate for a single comparison is 5% and seven treatments were compared? (Answer: 0.66) The appropriate procedure for testing the equality of several means is the analysis of variance, which is introduced in the context of a completely randomized design.

Completely Randomized Design The simplest randomized design for comparing several treatments is the completely randomized design (CRD). CRDs have $a \geq 2$ treatments to compare and N experimental units. Each treatment is applied to n_i ($i = 1, 2, \dots, a$) experimental units, where $n_1 + n_2 + \dots + n_a = N$. In order to conduct the experiment, the researcher randomly assigns treatments to the experimental units (design structure). Although the sizes of the a samples need not be identical, the power of the test is maximized

when $n_1 = n_2 = \dots = n_a$ for the a treatments. On each experimental unit, a response variable Y is measured. In Example 11.1 on page 493, Y represents the distance to the nearest foot required to stop a particular model of car traveling at 60 miles per hour using four different brands of tires. The CRD, when there is one factor with a levels (treatments) and no assumed relationships among the a levels, is called a **one-way treatment** structure. The layout for such a design is shown in Table 11.1.

Table 11.1: One-way design

Treatment					Totals	Means
1	Y_{11}	Y_{12}	\dots	Y_{1n_1}	$Y_{1\bullet}$	$\bar{Y}_{1\bullet} = \sum Y_{1j}/n_1$
2	Y_{21}	Y_{22}	\dots	Y_{2n_2}	$Y_{2\bullet}$	$\bar{Y}_{2\bullet} = \sum Y_{2j}/n_2$
\vdots	\vdots	\vdots		\vdots		\vdots
a	Y_{a1}	Y_{a2}	\dots	Y_{an_a}	$Y_{a\bullet}$	$\bar{Y}_{a\bullet} = \sum Y_{aj}/n_a$
					$Y_{\bullet\bullet}$	$\bar{Y}_{\bullet\bullet}$

Notation is critical, and the following conventions are used throughout the chapter. The sum of the observations in the i^{th} treatment group is $Y_{i\bullet} = \sum_{j=1}^{n_i} Y_{ij}$, and the mean of the observations in the i^{th} treatment group is $\bar{Y}_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} = \frac{Y_{i\bullet}}{n_i}$. The bar indicates a mean while the dot (\bullet) indicates that values have been added over the indicated subscript. The sum of all observations is $Y_{\bullet\bullet} = \sum_{i=1}^a \sum_{j=1}^{n_i} Y_{ij}$. The grand mean of all observations is denoted $\bar{Y}_{\bullet\bullet} = \frac{1}{N} \sum_{i=1}^a \sum_{j=1}^{n_i} Y_{ij} = \frac{Y_{\bullet\bullet}}{N}$.

To describe the observations, the linear statistical model

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij} \text{ for } i = 1, 2, \dots, a \text{ and } j = 1, 2, \dots, n_a \tag{11.1}$$

is used, where Y_{ij} is the j^{th} observation of the i^{th} treatment, μ is a parameter common to all treatments called the overall mean, τ_i is a parameter unique to the i^{th} treatment called the i^{th} treatment effect, and ϵ_{ij} is a random error component. For hypothesis testing, the model errors are assumed to be normally and independently distributed with mean zero and constant standard deviation ($NID(0, \sigma)$). The careful reader will realize that this implies the variance is assumed to be constant for all a treatments.

The model given in (11.1) can be used for two different scenarios with respect to the treatment effects. When the treatments are specifically chosen by the experimenter and there is no desire to extend the results to other treatments, the model is referred to as a **fixed effects model**. On the other hand, when the treatments are selected at random from a larger population of possible treatments and the experimenter would like to extend the conclusions of the experiment of all treatments in the population, the model is called a **random effects model**. What follows deals with the fixed effects model.

11.2 Fixed Effects Model

Although there are a means μ_i , one for each of the a treatments, model (11.1) uses $a + 1$ parameters (μ and the a τ_i s) to describe the a means. This implies that μ and τ_i are not uniquely determined. A frequently used solution is to impose the constraint

$\sum_{i=1}^a n_i \tau_i = 0$. When the n_i s are equal, the constraint can be written as $\sum_{i=1}^a \tau_i = 0$. Although other solutions to the overparameterized model exist, estimators for model (11.1) in this text will only consider the sum to zero constraint on the τ_i s as a solution for the overparameterized model. Different software packages often impose differing constraints on the overparameterized model, and the user should pay close attention to how the software computes estimates for the model. The natural and unbiased estimator for μ_i is $\bar{Y}_{i\bullet}$, the average of the observations in that treatment group. Likewise, the natural and unbiased estimator of μ is $\bar{Y}_{\bullet\bullet}$, the grand mean of all of the responses. Using either least squares or maximum likelihood to derive estimators of μ and τ_i for model (11.1) results in the aforementioned quantities. Verifying the least squares estimators of the parameters of model (11.1) (using the sum to zero constraint) as well as the maximum likelihood estimators of model (11.1) (using the sum to zero constraint) is left as an exercise for the reader.

Using maximum likelihood techniques, an estimator of σ^2 , $\hat{\sigma}^2$, is found to be

$$\sum_{i=1}^a \sum_{j=1}^{n_i} \frac{(Y_{ij} - \bar{Y}_{i\bullet})^2}{N}.$$

Unfortunately, the expected value of $\hat{\sigma}^2$ is $\frac{(N-a)\sigma^2}{N}$, which is a biased estimator of σ^2 . Since, $E(aX) = a \cdot E(x)$,

$$\frac{N}{N-a} \cdot \sum_{i=1}^a \sum_{j=1}^{n_i} \frac{(Y_{ij} - \bar{Y}_{i\bullet})^2}{N}$$

will yield an unbiased estimator of σ^2 :

$$\hat{\sigma}^2 = \sum_{i=1}^a \sum_{j=1}^{n_i} \frac{(Y_{ij} - \bar{Y}_{i\bullet})^2}{N-a}.$$

These facts are summarized in Table 11.2.

Table 11.2: Model, parameters, and estimators for fixed effects, one-way CRD

Model	Parameter	Estimator
$Y_{ij} = \mu + \tau_i + \epsilon_{ij}$	μ	$\bar{Y}_{\bullet\bullet}$
	μ_i	$\bar{Y}_{i\bullet}$
	τ_i	$\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}$
	ϵ_{ij}	$Y_{ij} - \bar{Y}_{i\bullet}$
	σ^2	$\sum_{i=1}^a \sum_{j=1}^{n_i} \frac{(Y_{ij} - \bar{Y}_{i\bullet})^2}{N-a}$

Although estimating the parameters for model (11.1) is important, the goal of the experimenter is generally to discern whether or not the a treatment means are equal, and if they are not equal, which treatments are better (for example, have a higher mean). Specifically, the null hypotheses of interest are

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_a \quad \text{versus} \quad H_1 : \mu_i \neq \mu_j \text{ for some } (i, j).$$

When the null hypothesis is true, all treatments have a common mean μ and an equivalent statement of the null hypothesis can be written in terms of the treatment effects as

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_a = 0 \quad \text{versus} \quad H_1 : \tau_i \neq 0 \text{ for some } i.$$

Consequently, testing the equality of treatment means is equivalent to testing that the treatment effects are all zero. As mentioned earlier, the appropriate procedure for testing the null hypothesis of equal treatment means is the analysis of variance, which is simply a decomposition of the total variability into its component parts, which is shown next.

11.3 Analysis of Variance (ANOVA) for the One-Way Fixed Effects Model

Consider the identity

$$Y_{ij} - \bar{Y}_{\bullet\bullet} = (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}) + (Y_{ij} - \bar{Y}_{i\bullet}) \quad (11.2)$$

which partitions the deviation of any observation from the grand mean into two parts. The first part, $(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})$, is the deviation of the i^{th} treatment mean from the grand mean. The second part is the deviation of the observation from the i^{th} treatment mean. Squaring and summing both sides of (11.2) produces

$$\begin{aligned} \sum_{i=1}^a \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{\bullet\bullet})^2 &= \sum_{i=1}^a \sum_{j=1}^{n_i} [(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}) + (Y_{ij} - \bar{Y}_{i\bullet})]^2 \\ &= \sum_{i=1}^a n_i (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 + \sum_{i=1}^a \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet})^2 \\ &\quad + 2 \sum_{i=1}^a \sum_{j=1}^{n_i} (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})(Y_{ij} - \bar{Y}_{i\bullet}) \end{aligned} \quad (11.3)$$

However, the cross product in (11.3) is zero since

$$\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet}) = Y_{i\bullet} - n_i \bar{Y}_{i\bullet} = Y_{i\bullet} - n_i \cdot \frac{Y_{i\bullet}}{n_i} = 0.$$

Consequently,

$$\sum_{i=1}^a \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^a n_i (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 + \sum_{i=1}^a \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet})^2 \quad (11.4)$$

which says the total variability in the data can be partitioned into two parts. The quantity $\sum_{i=1}^a n_i (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2$ measures the difference between the observed treatment means and the grand mean. Specifically, it is a measure of variability due to the treatments and is denoted $SS_{\text{Treatment}}$ (sum of squares due to treatments). The quantity $\sum_{i=1}^a \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet})^2$ measures the differences of observations within a treatment from the treatment mean, which must be due to error and is referred to as SS_{Error} (sum of squares due to error). The quantity on the left-hand side of the equals sign in (11.4) is called the total sum of squares corrected for the mean and is denoted SS_{Total} . The symbolic representation of (11.4) is

$$SS_{\text{Total}} = SS_{\text{Treatment}} + SS_{\text{Error}} \quad (11.5)$$

Since there are a total of $\sum_{i=1}^a n_i = N$ observations, SS_{Total} has $N - 1$ degrees of freedom. One degree of freedom is lost for estimating μ with the grand mean, $\bar{Y}_{\bullet\bullet}$. $SS_{\text{Treatment}}$

has $a - 1$ degrees of freedom since there are a treatment means and SS_{Error} has $N - a$ degrees of freedom. To adjust for the number of treatments, $SS_{\text{Treatment}}$ is divided by its degrees of freedom, $a - 1$. The resulting quantity is known as the **mean square treatment** $MS_{\text{Treatment}} = \frac{SS_{\text{Treatment}}}{df_{\text{Treatment}}}$ and is also called the between treatments error variance. In order to know whether the $MS_{\text{Treatment}}$ value is large, it is compared to an estimate of σ^2 , namely, MS_{Error} , where $MS_{\text{Error}} = \frac{SS_{\text{Error}}}{df_{\text{Error}}}$, which is also called the within treatments error variance. Note that MS_{Error} can be expressed as

$$\hat{\sigma}^2 = MS_{\text{Error}} = \frac{SS_{\text{Error}}}{df_{\text{Error}}} = \frac{\sum_{i=1}^a \left[\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet})^2 \right]}{df_{\text{Error}}} \quad (11.6)$$

If the term within the square braces is divided by its degrees of freedom ($n_i - 1$), it is easy to recognize that quantity as the sample variance for the i^{th} treatment:

$$S_i^2 = \sum_{j=1}^{n_i} \frac{(Y_{ij} - \bar{Y}_{i\bullet})^2}{n_i - 1}, \quad i = 1, 2, \dots, a \quad (11.7)$$

Combining the sample variances, a single estimate of the population variance emerges as

$$\begin{aligned} \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 + \dots + (n_a - 1)S_a^2}{(n_1 - 1) + (n_2 - 1) + \dots + (n_a - 1)} &= \frac{\sum_{i=1}^a \left[\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet})^2 \right]}{\sum_{i=1}^a (n_i - 1)} \\ &= \frac{SS_{\text{Error}}}{N - a} = MS_{\text{Error}} \end{aligned}$$

The pooled estimate of the variance from the two-sample t -test in Section 9.7.4 has now been generalized for a different samples.

If there are no differences among the a treatment means, $MS_{\text{Treatment}}$ is an unbiased estimate of σ^2 , and the ratio of $MS_{\text{Treatment}}/MS_{\text{Error}}$ will be close to 1. If differences actually exist among the a treatment means, then the ratio, $MS_{\text{Treatment}}/MS_{\text{Error}}$ should be larger than 1. In fact, it can be shown that

$$E(MS_{\text{Error}}) = \sigma^2 \quad \text{and} \quad E(MS_{\text{Treatment}}) = \sigma^2 + \sum_{i=1}^a \frac{n_i \tau_i^2}{a - 1}$$

implying that when H_0 is false, $E(MS_{\text{Treatment}}) > E(MS_{\text{Error}})$ since some $\tau_i \neq 0$. When H_0 is true, $\tau_i = 0$ for all i and $E(MS_{\text{Treatment}}) = E(MS_{\text{Error}}) = \sigma^2$. With a little effort, it can be shown that

$$\frac{MS_{\text{Error}}}{\sigma^2} \sim \frac{\chi_{df_{\text{Error}}}^2}{df_{\text{Error}}} = \frac{\chi_{N-a}^2}{N - a}$$

regardless of whether H_0 is true or not, and that

$$\frac{MS_{\text{Treatment}}}{\sigma^2} \sim \frac{\chi_{df_{\text{Treatment}}}^2}{df_{\text{Treatment}}} = \frac{\chi_{a-1}^2}{a - 1}$$

when H_0 is true independently of MS_{Error} . Consequently, using Definition 6.2 on page 238, when H_0 is true, the ratio $MS_{\text{Treatment}}/MS_{\text{Error}} \sim F_{a-1; N-a}$. Thus, H_0 is rejected in an α -level test if $F_{\text{Obs}} > f_{1-\alpha; a-1, N-a}$, where $F_{\text{Obs}} = MS_{\text{Treatment}}/MS_{\text{Error}}$. The `S` function `aov()` used with `summary()` returns a table similar to Table 11.3 on the facing page.

Example 11.1 ▷ **Tire ANOVA Table** ◁ Use the data frame `Tire` and compute the values for the ANOVA table using both the formulas and the `S` function `summary(aov())` to test the null hypothesis that all the tire treads have identical mean stopping distances versus the alternative hypothesis that there is at least one mean that is different.

Table 11.3: ANOVA table for one-way completely randomized design

Source of Variation (Source)	Degrees of Freedom (<i>df</i>)	Sum of Squares (<i>SS</i>)	Mean Square (<i>MS</i>)	<i>F</i>
Treatments	$a - 1$	$SS_{\text{Treatment}} = \sum_{i=1}^a n_i (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2$	$MS_{\text{Treatment}} = \frac{SS_{\text{Treatment}}}{a - 1}$	$\frac{MS_{\text{Treatment}}}{MS_{\text{Error}}}$
Error	$N - a$	$SS_{\text{Error}} = \sum_{i=1}^a \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet})^2$	$MS_{\text{Error}} = \frac{SS_{\text{Error}}}{N - a}$	
Total	$N - 1$	$SS_{\text{Total}} = \sum_{i=1}^a \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{\bullet\bullet})^2$		

Solution: The hypotheses being tested are

$$H_0 : \tau_i = 0 \text{ for all } i \quad \text{versus} \quad H_1 : \tau_i \neq 0 \text{ for some } i$$

$$\begin{aligned}
 SS_{\text{Treatment}} &= \sum_{i=1}^a n_i (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 \\
 &= 6 \cdot (379.6667 - 404.2083)^2 + 6 \cdot (405.1667 - 404.2083)^2 \\
 &\quad + 6 \cdot (421.6667 - 404.2083)^2 + 6 \cdot (410.3333 - 404.2083)^2 = 5673.12
 \end{aligned}$$

$$\begin{aligned}
 SS_{\text{Total}} &= \sum_{i=1}^a \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{\bullet\bullet})^2 \\
 &= (391 - 404.2083)^2 + (374 - 404.2083)^2 + (416 - 404.2083)^2 \\
 &\quad + \dots + (389 - 404.2083)^2 = 12771.96
 \end{aligned}$$

$$SS_{\text{Error}} = SS_{\text{Total}} - SS_{\text{Treatment}} = 12771.96 - 5673.12 = 7098.83$$

$$MS_{\text{Treatment}} = SS_{\text{Treatment}}/df_{\text{Treatment}} = 5673.12/3 = 1891.04$$

$$MS_{\text{Error}} = SS_{\text{Error}}/df_{\text{Error}} = 7098.83/20 = 354.94$$

$$F = MS_{\text{Treatment}}/MS_{\text{Error}} = 1891.04/354.94 = 5.33$$

The \wp -value for the test is $\mathbb{P}(F_{3,30} \geq 5.33) = 0.007$. Based on the small \wp -value, the null hypothesis of no tire tread effect ($\tau_i = 0$ for all i) is rejected. This suggests at least one tire tread effect is not zero. Thus, the question then becomes, “Which tire tread has the shortest mean stopping distance?” The statistical conclusion as well as the validity of any

Table 11.4: Tire ANOVA table

Source of Variation (Source)	Degrees of Freedom (<i>df</i>)	Sum of Squares (<i>SS</i>)	Mean Square (<i>MS</i>)	<i>F</i>
Treatments	$4 - 1 = 3$	5673.12	$\frac{5673.12}{3} = 1891.04$	$\frac{1891.04}{354.94} = 5.33$
Error	$24 - 4 = 20$	7098.83	$\frac{7098.83}{20} = 354.94$	
Total	$24 - 1 = 23$	12771.96		

multiple comparison procedures used to detect individual differences between tire treads assume the one-way model (11.1) is sound. Checking model assumptions for the one-way CRD is discussed in Section 11.5, followed by multiple comparison procedures in Section 11.7.

The values for the ANOVA Table can be computed with S by entering

```
> attach(Tire)
> TreatmentMean <- tapply(StopDist, tire, mean)
> TreatmentMean
      A      B      C      D
379.6667 405.1667 421.6667 410.3333
> a <- length(TreatmentMean)
> N <- length(StopDist)
> dft <- a - 1
> dfe <- N - a
> GrandMean <- mean(StopDist)
> GrandMean
[1] 404.2083
> SStreat <- 6*sum((TreatmentMean - GrandMean)^2)
> SStreat
[1] 5673.125
> SStotal <- sum((StopDist - GrandMean)^2)
> SStotal
[1] 12771.96
> SSerror <- SStotal - SStreat
> SSerror
[1] 7098.833
> MStreat <- SStreat/dft
> MStreat
[1] 1891.042
> MSerror <- SSerror/dfe
> MSerror
[1] 354.9417
> Fobs <- MStreat/MSerror
> Fobs
[1] 5.327753
> pvalue <- 1-pf(Fobs, 3, 20)
> pvalue
[1] 0.007315521
```

Or using the two functions `summary()` and `aov()` together returns

```
> summary(aov(StopDist~tire))
          Df Sum Sq Mean Sq F value    Pr(>F)
tire          3 5673.1  1891.0   5.3278 0.007316 **
Residuals    20 7098.8   354.9
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The treatment means and the grand mean can also be computed using the function `model.tables()`:

```
> model.tables(aov(StopDist~tire), type="means")
Tables of means
Grand mean

404.2083

tire
tire
  A     B     C     D
379.7 405.2 421.7 410.3
> detach(Tire)
```

11.4 Power and the Non-Central F Distribution

The concept of power and the non-central t -distribution for one-sample and two-sample problems was discussed in Section 9.7. In this section, computing power is extended to the $a \geq 2$ samples problem. Specifically, the problem of determining the required sample size to detect a given difference is addressed. Consider a slightly different but equivalent expression for the $MS_{\text{Treatment}}$ and MS_{Error} given in Table 11.3 on page 499 when $a = 2$.

$$F = \frac{MS_{\text{Treatment}}}{MS_{\text{Error}}} = \frac{\sum_{i=1}^2 \sum_{j=1}^{n_i} (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2}{df_{\text{Treatment}}} \equiv \frac{(\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet})^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1}}{\frac{\sum_{i=1}^2 \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet})^2}{df_{\text{Error}}}} \quad (11.8)$$

The reader should verify that the right-hand side of (11.8) is the same as the expression on the left of the \equiv . Two facts that should be kept in mind during the verification are $\sum_{j=1}^{n_i} Y_{ij} = \bar{Y}_{i\bullet} n_i$ for $i = 1, 2$ and $(n_1 + n_2) \bar{Y}_{\bullet\bullet} = \sum_{i=1}^2 \sum_{j=1}^{n_i} Y_{ij}$. Rewriting the right side of (11.8) gives

$$F = \frac{(\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet})^2}{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \left[\frac{(\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet})}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right]^2 = [t]^2. \quad (11.9)$$

One verifies that the pooled t -test from Section 9.7 is simply a special case of the F -test used in ANOVA when $a = 2$. It is important to emphasize that the equivalence of the pooled

t -test and the F -test used in ANOVA for $a = 2$ groups applies only to the non-directional hypothesis $H_1 : \mu_1 \neq \mu_2$ because

$$[t_{1-\alpha/2; df_{\text{Error}}}]^2 = f_{1-\alpha; 1, df_{\text{Error}}}, \text{ but} \quad (11.10)$$

$[t_{1-\alpha; df_{\text{Error}}}]^2 \neq f_{1-\alpha; 1, df_{\text{Error}}}$, as would be required for a directional hypothesis. In Section 9.7, the non-centrality parameter γ for the pooled t -test was defined as

$$\gamma = \frac{\mu_1(X, Y) - \mu_0(X, Y)}{\sigma_{\bar{X} - \bar{Y}}}.$$

An equivalent expression for defining the non-centrality parameter is

$$\gamma = \frac{(\mu_1 - \mu_2) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{-1/2}}{\sigma}. \quad (11.11)$$

One should take note of the similarities between

$$t = \frac{(\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{-1/2}}{S_p} \quad (11.12)$$

and (11.11). Specifically, the quantity in (11.12) is used to measure the statistical differences between the **sample** means. In a similar fashion, (11.11) is used to measure the statistical differences between the population means. Rewriting (11.12) and (11.11), one notes

$$F = t^2 = \frac{(\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet})^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{-1}}{S_p^2} = \frac{MS_{\text{Treatment}}}{MS_{\text{Error}}}$$

and

$$\lambda = \gamma^2 = \frac{(\mu_1 - \mu_2)^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{-1}}{\sigma^2} = \frac{SS_{\text{Hypothesis}}(\text{population})}{\sigma^2}$$

where $SS_{\text{Hypothesis}}(\text{population})$ is the sum of squares for treatments obtained by replacing $\bar{Y}_{1\bullet}$ with μ_1 , $\bar{Y}_{2\bullet}$ with μ_2 , and $\bar{Y}_{\bullet\bullet}$ with $\frac{n_1\mu_1 + n_2\mu_2}{n_1 + n_2}$. By defining the non-centrality parameter λ as the ratio of $SS_{\text{Hypothesis}}(\text{population})$ to σ^2 , it becomes easy to calculate λ using statistical software. The $SS_{\text{Hypothesis}}(\text{population})$ will always be the sum of squares formula for the H_0 being tested, thus this method of computing λ extends to whatever hypothesis the user would like to test. It is not limited merely to the equality of treatment means. For any completely randomized design, $SS_{\text{Hypothesis}}(\text{population}) = \sum_{i=1}^a n_i (\mu_{i\bullet} - \bar{\mu}_{\bullet\bullet})^2$, where $\bar{\mu}_{\bullet\bullet} = (\sum_{i=1}^a n_i \mu_{i\bullet}) / (\sum_{i=1}^a n_i)$. Recall that power is the probability that the null hypothesis will be rejected when it is false. In this case,

$$\text{Power}(\lambda) = \mathbb{P}[F_{a-1; N-a, \lambda} > f_{1-\alpha; a-1; N-a; \lambda=0}]. \quad (11.13)$$

$\text{Power}(\lambda)$ is maximized when all a groups have an equal number of observations; however, using $SS_{\text{Hypothesis}}$ to compute the non-centrality parameter adjusts for experiments with different sample sizes. R has the function `power.anova.test()`, which can be used to determine the sample size for the a samples when resources are allocated such that each group has the same size.

Example 11.2 ▷ *Tires' Stopping Distance* ◁ Suppose the tire manufacturer believes the true mean stopping distance for tread patterns A, B, C, and D to be 390, 405, 415, and 410 feet, respectively, with a common standard deviation that could be as high as 20 feet or as small as 10 feet. Assume sets of tires are put on the car (a single car is used for all tests to reduce variability) in random order.

- Suppose the manufacturer wants to test $H_0 : \mu_B - \mu_A = 0$ versus $H_1 : \mu_B - \mu_A > 0$ using $\alpha = 0.05$, assuming $\sigma = 10$. Determine the power of the test if six sets of tires with each tread are available.
- Determine the probability that differences among the means will be detected using $\alpha = 0.05$ assuming $\sigma = 20$ feet if six sets of tires with each tread are available. Simulate the non-central F distribution and compute the power by simulation. How does the simulation compare to the theoretical answer?
- Determine the probability that differences among the means will be detected using $\alpha = 0.05$ if six sets of tires with each tread are available and assuming $\sigma = 10$ feet.
- Assuming the stopping distance standard deviation for all tire sets is $\sigma = 20$ feet, what is the minimum number of tire sets that need to be used to ensure the probability of detecting tire tread differences is at least 80%?
- Given 6 sets of tires with tread A, 6 sets of tires with tread B, 12 sets of tires with tread C, and 12 sets of tires with tread D, what is the probability of detecting tire tread differences if the true stopping standard deviation for all tire tread sets is $\sigma = 14$ feet?

Solution: The answers are as follows:

(a)

$$\lambda = \frac{SS_{\text{Hypothesis}}}{\sigma^2} = \frac{\sum_{i=1}^a n_i (\mu_{i\bullet} - \bar{\mu}_{\bullet\bullet})^2}{\sigma^2} = \frac{6 \cdot (405 - 397.5)^2 + 6 \cdot (390 - 397.5)^2}{10^2} = 6.75$$

$\text{Power}(\gamma = \sqrt{\lambda} = 2.598) = \mathbb{P}(t_{10; \gamma=2.598} > t_{0.95; 10} = 1.81) = 0.78$. This can be computed with R using the following code:

```
> MEANS <- c(405, 390)
> a <- length(MEANS)
> n <- 6
> N <- a*n
> dfe <- N - a
> SD <- 10
> alpha <- .05
> Y <- rep(MEANS, rep(n, a))
> treat <- factor(rep(1:a, rep(n, a)))
> SStreat <- summary(aov(Y~treat))[[1]][1, 2]
> lambda <- SStreat/SD^2
> Gamma <- sqrt(lambda)
> Gamma
[1] 2.598076
> cv <- qt(1 - alpha, dfe)
> Power <- 1 - pt(cv, dfe, ncp=Gamma)
> Power
[1] 0.7798662
```

Using the function `power.t.test()` gives

```
> power.t.test(n=6, delta=15, sd=10, alternative="one.sided")
```

```
Two-sample t test power calculation
```

```

      n = 6
    delta = 15
      sd = 10
sig.level = 0.05
  power = 0.7798662
alternative = one.sided

```

NOTE: `n` is number in *each* group

Note that the alternative hypothesis is directional and the F distribution cannot be used to answer the question. The answer is obtained with a non-central t -distribution. A graphical representation of the power is given in Figure 11.4.

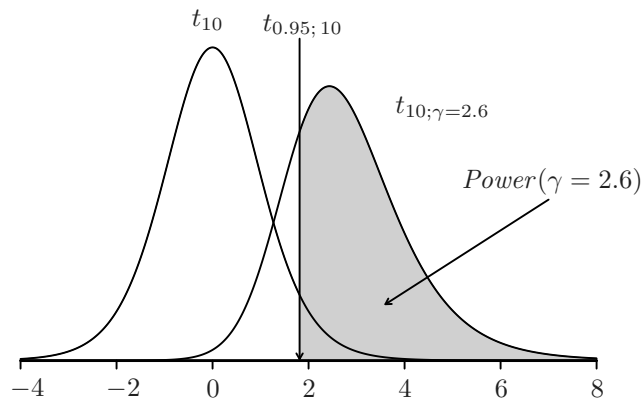


FIGURE 11.4: Power for the directional alternative hypothesis $H_1 : \mu_B - \mu_A > 0$ when $\gamma = 2.6$ at the $\alpha = 0.05$ level

(b) The following R code is used to simulate the non-central F distribution:

```

> ### Sampling Distribution of MST/MSE
> set.seed(10)
> sims <- 10000 # number of simulations
> n1 <- 6; n2 <- 6; n3 <- 6; n4 <- 6
> a <- 4 # number of treatments
> N <- n1+n2+n3+n4
> df.treat <- a - 1 # dof treatment
> df.error <- N - a # dof error

```

```

> alpha <- 0.05          # alpha level
> ### Normal Distribution
> mu1 <- 390; sig1 <- 20 # pop1 mean and sigma
> mu2 <- 405; sig2 <- 20 # pop2 mean and sigma
> mu3 <- 415; sig3 <- 20 # pop3 mean and sigma
> mu4 <- 410; sig4 <- 20 # pop4 mean and sigma
> t1 <- matrix(rnorm(sims*n1, mu1, sig1), nrow=sims, byrow=TRUE)
> t2 <- matrix(rnorm(sims*n2, mu2, sig2), nrow=sims, byrow=TRUE)
> t3 <- matrix(rnorm(sims*n3, mu3, sig3), nrow=sims, byrow=TRUE)
> t4 <- matrix(rnorm(sims*n4, mu4, sig4), nrow=sims, byrow=TRUE)
> MUS <- c(mu1, mu2, mu3, mu4)
> MUB <- mean(MUS)
> lambda <- (n1*(mu1-MUB)^2 + n2*(mu2-MUB)^2 + n3*(mu3-MUB)^2 +
+ n4*(mu4-MUB)^2)/(sig1^2)
> mt1 <- apply(t1, 1, mean)
> mt2 <- apply(t2, 1, mean)
> mt3 <- apply(t3, 1, mean)
> mt4 <- apply(t4, 1, mean)
> #####
> mmean <- cbind(mt1, mt2, mt3, mt4)
> TT <- cbind(t1, t2, t3, t4)
> gm <- apply(TT, 1, mean)
> SStreat <- n1*((mt1 - gm)^2) + n2*((mt2 - gm)^2) + n3*((mt3 - gm)^2) +
+ n4*((mt4-gm)^2)
> JU2 <- (TT - gm)^2
> SStotal <- apply(JU2, 1, sum)
> SSerror <- SStotal - SStreat
> Fobs <- (SStreat/df.treat)/(SSerror/df.error)
> q995 <- quantile(Fobs, .995)
> #####
> hist(Fobs, col="pink", prob=TRUE, breaks="Scott", main="",
+ xlim=c(0, q995))
> title(main="Simulated Sampling Distribution")
> curve(df(x, df.treat, df.error, lambda), 0, q995, col="red",
+ add=TRUE, lwd=3) # only R
> val <- c(.80, .85, .90, .95, .99)
> Theoretical <- qf(val, df.treat, df.error, lambda)
> Simulated <- quantile(Fobs, val)
> SimSigLev <- c( sum(Fobs>Theoretical[1])/sims,
+ sum(Fobs>Theoretical[2])/sims,
+ sum(Fobs>Theoretical[3])/sims,
+ sum(Fobs>Theoretical[4])/sims,
+ sum(Fobs>Theoretical[5])/sims )
> TheSigLev <- 1 - val
> ANS <- rbind(Theoretical, Simulated, TheSigLev, SimSigLev)
> round(ANS, 4)

```

	80%	85%	90%	95%	99%
Theoretical	4.5080	5.1095	5.9577	7.4323	11.0979
Simulated	4.6006	5.1574	6.0031	7.4895	10.7363
TheSigLev	0.2000	0.1500	0.1000	0.0500	0.0100
SimSigLev	0.2094	0.1527	0.1024	0.0508	0.0090

```
> #####
> Simulated.Power <- sum(Fobs > qf(1 - alpha, df.treat, df.error))/sims
> Simulated.Power
[1] 0.3984
```

The histogram of the simulated non-central F distribution is found in Figure 11.5.

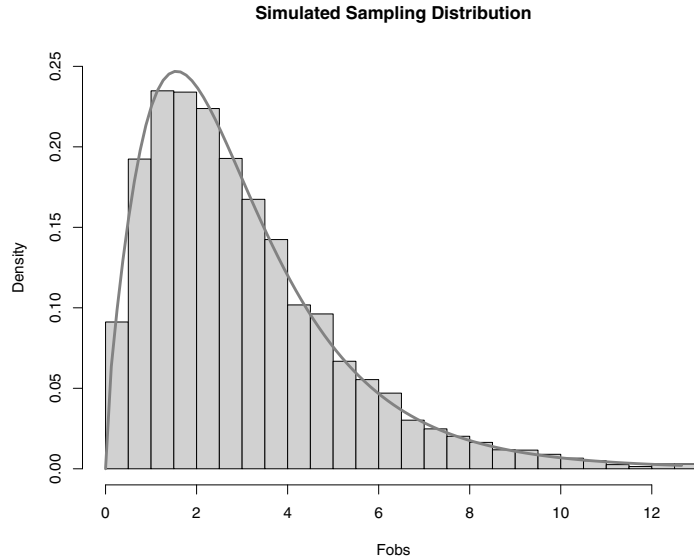


FIGURE 11.5: Histogram of simulated $F_{3, 20; \lambda=5.25}$ superimposed by the theoretical distribution

$$\begin{aligned}\lambda &= \frac{SS_{\text{Hypothesis}}}{\sigma^2} = \frac{\sum_{i=1}^a n_i (\mu_{i\bullet} - \bar{\mu}_{\bullet\bullet})^2}{\sigma^2} \\ &= \frac{6 \cdot (390 - 405)^2 + 6 \cdot (405 - 405)^2 + 6 \cdot (415 - 405)^2 + 6 \cdot (410 - 405)^2}{20^2} = \frac{2100}{20^2} = 5.25\end{aligned}$$

$$\text{Power}(\lambda = 5.25) = \mathbb{P}(F_{3, 20; \lambda=5.25} > f_{0.95; 3, 20} = 3.098) = 0.386.$$

A graphical representation of the central and non-central F distributions along with a shaded region for the power at $\lambda = 5.25$ is shown in Figure 11.6 on the facing page.

Using S:

```
> MEANS <- c(390, 405, 415, 410)
> a <- length(MEANS)
> n <- 6
> N <- a*n
> SD <- 20
> alpha <- .05
> Y <- rep(MEANS, rep(n, a))
```

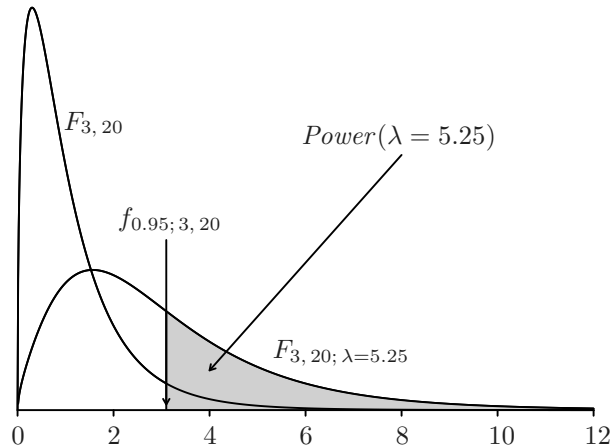


FIGURE 11.6: Power for detecting treatment differences when $\lambda = 5.25$ at the $\alpha = 0.05$ level

```
> treat <- factor(rep(1:a, rep(n, a)))
> SStreat <- summary(aov(Y~treat))[[1]][1, 2] # For R
> # SStreat <- summary(aov(Y~treat))[1, 2]    # For S-PLUS
> lambda <- SStreat/SD^2
> lambda
[1] 5.25
> cv <- qf(1-alpha, a - 1, N - a)
> Power <- 1-pf(cv, a - 1, N - a, ncp=lambda)
> Power
[1] 0.3862415
```

Since sample sizes are equal in the a groups, the R function `power.anova.test()` can be used to solve the problem:

```
> power.anova.test(groups=a, n=n, between.var=var(MEANS), within.var=SD^2)
```

Balanced one-way analysis of variance power calculation

```
groups = 4
n = 6
between.var = 116.6667
within.var = 400
sig.level = 0.05
power = 0.3862415
```

NOTE: n is number in each group

(c)

$$\begin{aligned}
 \lambda &= \frac{SS_{Hypothesis}}{\sigma^2} \\
 &= \frac{\sum_{i=1}^a n_i (\mu_{i\bullet} - \bar{\mu}_{\bullet\bullet})^2}{\sigma^2} \\
 &= \frac{6 \cdot (390 - 405)^2 + 6 \cdot (405 - 405)^2 + 6 \cdot (415 - 405)^2 + 6 \cdot (410 - 405)^2}{20} \\
 &= \frac{2100}{10^2} = 21
 \end{aligned}$$

$$Power(\lambda = 21) = \mathbb{P}(F_{3, 20; \lambda=21} > f_{0.95; 3, 20} = 3.098) = 0.95.$$

Using S:

```

> MEANS <- c(390, 405, 415, 410)
> a <- length(MEANS)
> n <- 6
> N <- a*n
> SD <- 10
> alpha <- .05
> Y <- rep(MEANS, rep(n, a))
> treat <- factor(rep(1:a, rep(n, a)))
> SStreat <- summary(aov(Y~treat))[[1]][1, 2] # For R
> # SStreat <- summary(aov(Y~treat))[1, 2] # For S-Plus
> lambda <- SStreat/SD^2
> lambda
[1] 21
> cv <- qf(1-alpha, a - 1, N - a)
> Power <- 1-pf(cv, a - 1, N - a, ncp=lambda)
> Power
[1] 0.9501649

```

Again, because treatment groups have equal n s, an answer is possible using R's function `power.anova.test()`:

```

> SD <- 10
> power.anova.test(groups=a, n=n, between.var=var(MEANS), within.var=SD^2)

```

Balanced one-way analysis of variance power calculation

```

groups = 4
n = 6
between.var = 116.6667
within.var = 100
sig.level = 0.05
power = 0.9501649

```

NOTE: n is number in each group

(d) Since λ is a function of sample size, one solution is to find n such that $\mathbb{P}(F_{a-1, a-n-a, \lambda} > f_{0.95, a-1, a-n-a}) \geq 0.80$. The following code uses a loop to find the value of n such that

the power is at least 80%. Power is maximized with a total of N sets when each of the a treatments receives n sets of tires such that $N = a \cdot n$. That is, power is maximized with equal treatment sizes.

```
> SD <- 20
> Powerr <- 0
> nr <- 1
> MEANS <- c(390, 405, 415, 410)
> a <- length(MEANS)
> while(Powerr < .80)
+ {
+   nr <- nr + 1
+   Nr <- a*nr
+   alpha <- .05
+   Yr <- rep(MEANS, rep(nr, a))
+   treatr <- factor(rep(1:a, rep(nr, a)))
+   SStreatr <- summary(aov(Yr~treatr))[[1]][1, 2] # R
+   # SStreatr <- summary(aov(Yr~treatr))[1, 2] # S-PLUS
+   lambdar <- SStreatr/SD^2
+   cvr <- qf(1 - alpha, a - 1, Nr - a)
+   Powerr <- 1 - pf(cvr, a - 1, Nr - a, ncp=lambdar)
+ }
> c(nr, lambdar, Powerr)
[1] 14.0000000 12.2500000 0.8176811
```

From the output, note that when $n = 14$, $\lambda = 12.25$, which returns a power of 81.7%. Since the problem permits equal n per treatment group, the R function `power.anova.test()` can also be used:

```
> power.anova.test(groups=4, power=.80,
+ between.var=var(MEANS), within.var=20^2)
```

Balanced one-way analysis of variance power calculation

```
groups = 4
n = 13.47806
between.var = 116.6667
within.var = 400
sig.level = 0.05
power = 0.8
```

NOTE: n is number in each group

```
> nrf <- ceiling(power.anova.test(groups=4, power=.80,
+ between.var=var(MEANS), within.var=20^2)$n)
> nrf
[1] 14
```

(e) Using the following code, the power is computed as 0.83:

```
> MEANS <- c(390, 405, 415, 410)
> SD <- 14
```



```

> a <- length(MEANS)
> n1=6; n2=6; n3=12; n4=12;
> N <- n1 + n2 + n3 + n4
> alpha <- .05
> Y <- rep(MEANS, c(n1, n2, n3, n4))
> treat <- factor(rep(1:a, c(n1, n2, n3, n4)))
> SStreat <- summary(aov(Y~treat))[[1]][1, 2] # R
> # SStreat <- summary(aov(Y~treat))[1, 2] # S-PLUS
> lambda <- SStreat/SD^2
> lambda
[1] 13.39286
> cv <- qf(1 - alpha, a - 1, N - a)
> Power <- 1 - pf(cv, a - 1, N - a, ncp=lambda)
> Power
[1] 0.8349338

```

■

11.5 Checking Assumptions

The values in the ANOVA table and the subsequent inferences made from those values are based on the assumption that the data follow the model

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad (11.14)$$

where the τ_i s are fixed but unknown numbers and the ε_{ij} s are independent normals with a mean of zero and constant variance. Consequently, the three basic assumptions concerning the errors:

- 1) independence,
- 2) normal distribution, and
- 3) constant variance

should be investigated. Since the actual errors are unknown quantities, they will never be observed; however, it is possible to use estimates (or predictors) of the errors, the residuals. Recall from Chapter 2 that a residual is the difference between what is observed and what is predicted ($\hat{\varepsilon}_{ij} = Y_{ij} - \hat{Y}_{ij}$). For model (11.14), $\hat{Y}_{ij} = \bar{Y}_{\bullet\bullet} + \bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet} = \bar{Y}_{i\bullet}$. While (11.14) may be a reasonable approximation to some real-life phenomena, real-life data are never exactly normal. The real question is whether the assumptions have been violated to such an extent that the inferences based on the particular model in question would be invalidated. Although a few formal tests are presented, most of the material that follows deals with visual diagnostics for the three basic assumptions concerning errors.

11.5.1 Checking for Independence of Errors

The most important assumption for (11.14) to be valid and the most challenging assumption to correct if it fails is the assumption of independence. The material in this text will not address how to deal with dependent data, which is the topic of a more advanced

course. One of the easier dependencies to detect is a dependence in time. When values are either very similar (positive dependence) or very different (negative dependence) to each other in time, the assumption of independence becomes untenable. An easy way visually to inspect data for dependence is to plot the residuals on the vertical axis versus a time sequence on the horizontal axis. Naturally, if there is no time component to the data, this graph will not reveal any useful information not found in other residual plots.

It is often helpful to standardize the residuals so they have unit variance. Many books define standardized residuals as

$$r_{ij} = \frac{\hat{\varepsilon}_{ij}}{\sqrt{MS_{\text{Error}}}},$$

however, the standard deviation of the ij^{th} residual is actually $\sigma \cdot \sqrt{1 - h_{ii}}$, where the h_{ii} s are the diagonal elements of the hat matrix (discussed in more detail in Chapter 12: Regression). For model (11.14), the h_{ii} values are simply $1/n_i$. By estimating σ with the $\sqrt{MS_{\text{Error}}}$, the standardized residuals (r_{ij}) are computed as

$$r_{ij} = \frac{\hat{\varepsilon}_{ij}}{\sqrt{\widehat{\text{Var}}(\hat{\varepsilon}_{ij})}} = \frac{\hat{\varepsilon}_{ij}}{\sqrt{MS_{\text{Error}} \cdot \sqrt{1 - h_{ii}}}} \quad (11.15)$$

The function `stdres()` in the `MASS` package computes standardized residuals according to (11.15).

Modifications to the following code used with the `Tire` data set from the motivational problem (Example 11.1) at the beginning of the chapter can be used to help the user assess the assumption of independence among the errors. Based on Figure 11.7 on the next page, no discernible pattern is seen that might threaten the assumption of independent errors.

```
> attach(Tire)
> par(pty="s")
> mod.aov <- aov(StopDist~tire)
> library(MASS)
> r <- stdres(mod.aov)
> n <- length(StopDist)
> plot(1:n, r, ylab="Standardized Residual", xlab="Ordered Value")
> detach(Tire)
```

11.5.2 Checking for Normality of Errors

The quantile-quantile plot is a graphical procedure for assessing normality. The quantile-quantile plot can be performed on either the residuals or the standardized residuals. If standardized residuals are used, the plotted observations should follow a straight line with an intercept of zero and a slope of one. Reading quantile-quantile plots, especially when the total number of residuals ($N = \sum_{i=1}^a n_i$) is small, requires a high degree of skill. A formal test of normality can be obtained with the function `shapiro.test()`. Modifications to the following code used with the `Tire` data set from the motivational problem (Example 11.1) can be used to help the user assess the normality of errors assumption. Figure 11.8 on page 513 shows a quantile-quantile plot of the standardized residuals with a superimposed line with a zero intercept and a slope of one indicating the assumption of normal errors is reasonable. The ϕ -value (0.7584) from the Shapiro-Wilk normality test provides further corroboration that the normality assumption of the errors is reasonable.

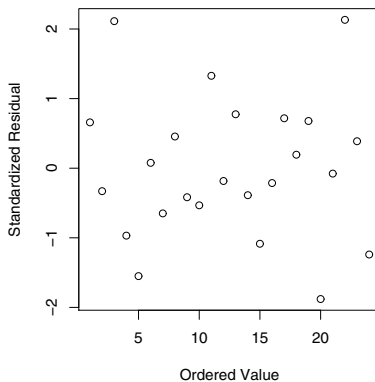


FIGURE 11.7: Standardized residuals versus order for `mod.aov` using the `Tire` data set

```
> attach(Tire)
> mod.aov <- aov(StopDist~tire)
> library(MASS)
> r <- stdres(mod.aov)
> par(pty="s")
> qqnorm(r)
> abline(a=0, b=1)
> shapiro.test(r)
```

Shapiro-Wilk normality test

```
data: r
W = 0.9737, p-value = 0.7584
```

```
> detach(Tire)
```

11.5.3 Checking for Constant Variance

Many formal tests for equality of variance exist. Most of these tests are very sensitive to normality assumptions and will not give reliable results if normality is violated. As with independence and normality of errors assumptions, the constant variance assumption should be checked with graphical procedures. Specifically, to assess constant variance, a plot of the residuals ($\hat{\varepsilon}_{ij}$) or the standardized residuals r_{ij} on the vertical axis should be plotted against the fitted values (\hat{Y}_{ij}) on the horizontal axis. Recall that for (11.14), the fitted values are simply $\bar{Y}_{i\bullet}$. This plot will look like several vertical stripes of points, one for each treatment group. If the variance is constant, the vertical lengths of the stripes for each of the i groups will be similar. If one insists on testing equality of variance, a modified version of Levene's test is recommended. Specifically, compute the quantity $Z_{ij} = |Y_{ij} - \tilde{Y}_{i\bullet}|$, the absolute deviations from the group medians. Treat the Z_{ij} values as the data, and use the standard ANOVA formulas presented in Table 11.2 on page 496 on the Z_{ij} values. A significant finding with the standard F -test on the Z_{ij} values indicates non-constant variance. This

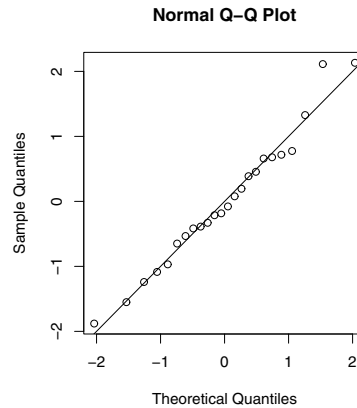


FIGURE 11.8: Quantile-quantile plot of the standardized residuals with a superimposed line with a zero intercept and a slope of one for the model `mod.aov` using the `Tire` data frame

particular modification to Levene's test, which uses the absolute deviations from the group medians, is relatively insensitive to non-normality and is easily implemented with `S`. It is also a preprogrammed function `levene.test()` in the `car` package. Modifications to the following code used with the `Tire` data set from Example 11.1 can be used to help the user assess the homogeneity of variance with respect to the errors assumption. Figure 11.9 on the following page shows a plot of the standardized residuals versus the fitted values of (11.14), indicating that there are no serious departures in homogeneity of variance. The fitted values (\hat{Y}_{ij}) of an `aov` object can be obtained by using the `fitted()` on an `aov` object. The p -value (0.4224) from the modified Levene test provides further corroboration that the homogeneity of variance assumption of the errors is reasonable.

```
> attach(Tire)
> mod.aov <- aov(StopDist~tire)
> library(MASS)
> r <- stdres(mod.aov)
> tm <- fitted(mod.aov)
> plot(tm, r, xlab="Fitted Value", ylab="Standardized Residual")
> med <- tapply(StopDist, tire, median)
> ZIJ <- abs(StopDist - med[tire])
> summary(aov(ZIJ~tire))
      Df Sum Sq Mean Sq F value Pr(>F)
tire    3  388.79  129.60  0.9789 0.4224
Residuals 20 2647.83  132.39
> checking.plots(mod.aov)
> detach(Tire)
```

The function `checking.plots()` from the `PASWR` package creates the three graphs discussed in Sections 11.5.1, 11.5.2, and 11.5.3 that assess independence, normality, and constant variance, respectively. The graphs from using `checking.plots()` with `mod.aov` are shown in Figure 11.10 on the next page.

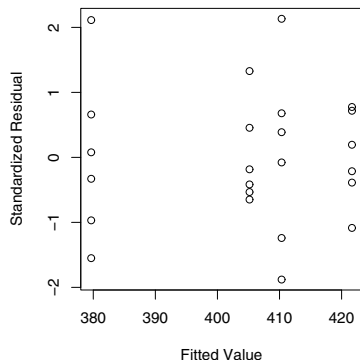


FIGURE 11.9: Plot of the standardized residuals versus the fitted values for `mod.aov` using the `Tire` data set

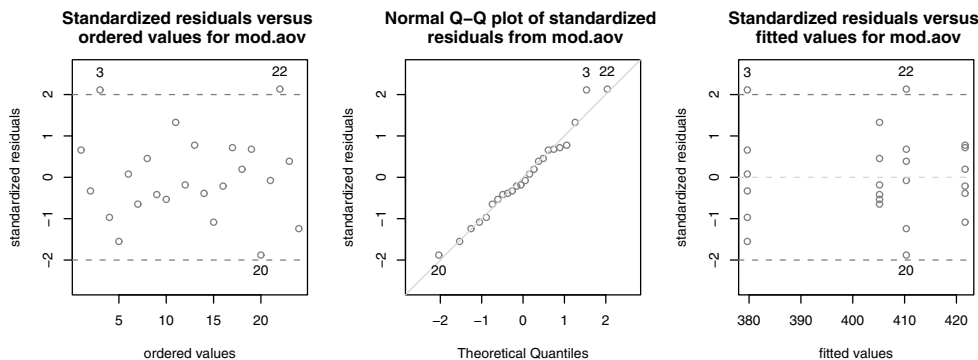


FIGURE 11.10: Graphs to assess independence, normality, and constant variance, respectively, created with `checking.plots(mod.aov)` using the data frame `Tire`

11.6 Fixing Problems

When diagnostics indicate that the assumptions for a particular model are not satisfied, either the data must be modified or the method of analysis must be changed to be less sensitive to the assumptions. The three assumptions for error terms of (11.14) are that they are 1) independent, 2) have a normal distribution, and 3) have homogeneity of variance. Working with dependent errors is quite challenging and will not be discussed other than to say that proper randomization should always be used in the collection of data to reduce the possibility of dependence among errors. In the event an analysis indicates dependent errors, the original design should be re-evaluated. The normal errors assumption can often be violated without affecting the estimation and inferences associated with the chosen model provided the errors' departures from normality are not severe. Non-constant variance in contrast to the normality assumption will impact estimation and inferences associated with

the chosen model and needs to be evaluated closely. The balance of fixing problems will center on how to deal with 1) non-normal errors and 2) non-constant variance.

11.6.1 Non-Normality

When a quantile-quantile plot of the residuals indicates skewness (typically to the right) a transformation on the response variable will often alleviate the problem of non-normal errors. Finding a meaningful and appropriate transformation is often challenging. One technique that searches computationally for an appropriate transformation of the response variable that directly addresses normality is the Box-Cox method. The Box-Cox method estimates the parameter λ for the transformation $Y' = Y^\lambda$, where

$$Y' = \begin{cases} \frac{Y^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0, \\ \ln Y & \text{for } \lambda = 0, \end{cases} \quad (11.16)$$

by the method of maximum likelihood. Figure 11.11 shows transformations in common use. The function `boxcox()` of the MASS package produces a plot of the log-likelihood against the

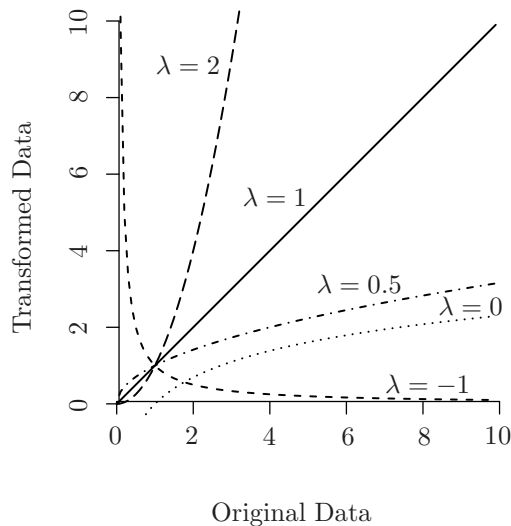


FIGURE 11.11: Transformations in common use with the Box-Cox method: The long dashed line shows data transformed by squaring; the solid, by doing nothing; the dot-dashed, by taking the square root; the dotted, by taking the natural log; and the dashed, by taking the reciprocal

transformation parameter λ for a particular model. By default, the range of λ is from -2 to 2 . However, once the value of λ that maximizes the log-likelihood is known, the range of the plot in `boxcox()` can be tightened to highlight the area where the function is maximized with the argument `lambda=`. For more details, see the `boxcox()` help file. The `boxcox()` function is generally used to approximate an appropriate transformation. The value of λ

that maximizes the log-likelihood function may turn out to be 0.53; but if there is a possible explanation for taking the square root of the response, the transformation applied should be $\lambda = 0.5$ and not the value that maximizes the log-likelihood function.

Observations that do not fit the pattern of the rest of the data in the quantile-quantile plot (outliers) can distort an analysis, and one should consider removing the outlier(s) and performing the analysis without the offending point(s). Oftentimes, outliers are simply poorly transcribed experimental results such as an incorrectly placed decimal or a misplaced label. However, just because a value is an outlier does not mean it should be eliminated from the data; rather, outliers imply that the model being used is incorrect. Does this suggest that if the values in a quantile-quantile plot are not exactly linear, then there are problems? Fortunately not! With equal treatment sizes, the F -test used with ANOVA is quite robust to non-normal errors when the homogeneity of variance assumption is satisfied. The reader should perform their own simulations to verify that sampling distribution for $MS_{\text{Treatment}}/MS_{\text{Error}}$ when sampling from non-normal distributions is quite close to the F distribution. Unfortunately, subsequent inference on individual parameters using one-sided confidence intervals is sensitive to the normality assumption and can result in poor conclusions when the errors do not follow a normal distribution.

11.6.2 Non-Constant Variance

Non-constant variance is typically fixed by transforming the response variable. The Box-Cox method discussed to fix the problem of non-normal errors will oftentimes alleviate both the problem of unequal variances as well as non-normal errors. The implications for the F -test when the variances among the a groups are different depends to a large extent on whether the groups have equal sample sizes. When the a groups have equal sample sizes, unequal variance only slightly alters the p -value for an F -test. The situation is very different, however, when sample sizes among the a groups are unequal. When larger variances are associated with the smaller sample sizes, the F -test will be conservative, and when the larger variances are associated with smaller sample sizes, the F -test is liberal. Welch (1951) derived a method of testing several means that does not require the assumption of equal variance and is implemented in R using the function `oneway.test()`. Welch's statistic (W) for testing several means is defined as

$$W = \frac{\sum_{i=1}^a w_i (\bar{Y}_i - \tilde{Y})^2 / (a-1)}{\left[1 + \frac{2}{3}(a-2)\Lambda\right]} \sim F_{a-1; 1/\Lambda} \quad (11.17)$$

where

$$w_i = \frac{n_i}{s_i^2}, \quad \tilde{Y} = \frac{\sum_{i=1}^a w_i \bar{Y}_i}{\sum_{i=1}^a w_i}, \quad \text{and} \quad \Lambda = \frac{3 \sum_{i=1}^a \left\{ \left[1 - \left(w_i / \sum_{i=1}^a w_i \right) \right]^2 / (n_i - 1) \right\}}{a^2 - 1}.$$

Example 11.3 ▷ *Fat Cats* ◁ In a weight loss study on obese cats, overweight cats were randomly assigned to one of three groups and boarded in a kennel. In each of the three groups, the cats' total caloric intake was strictly controlled (1 cup of generic cat food) and monitored for 10 days. The difference between the groups was that group A was given $1/4$ of a cup of cat food every 6 hours, group B was given $1/3$ a cup of cat food every 8 hours, and group C was given $1/2$ a cup of cat food every 12 hours. The weights of the cats at the beginning and end of the study were recorded and the differences in weights (grams) are stored in the variable `Weight` of the data frame `FCD`. Are there mean weight differences among the three treatments?

Solution: The hypothesis of interest is $H_0 : \tau_i = 0$ for all i versus $H_1 : \tau_i \neq 0$ for some i given the model $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$, where $\varepsilon_{ij} \sim N(0, \sigma)$. To see if the assumption of NID errors is reasonable, the function `checking.plots()` is applied to the model `FCD.aov <- aov(Weight~Diet)` and the graphical output is displayed in Figure 11.12. The assumption of equal variance seems tenuous. The increasing variance as the mean increases is seen in the standardized residuals versus fitted graph (third graph). By using the function `boxcox()` from the MASS package applied to `FCD.aov`, a log transformation is suggested (see Figure 11.13 on the following page). However, the log transformation does not fix the unequal variance assumption (see Figure 11.14 on the next page). It is interesting to point out in this particular case that, despite the logarithmic transformation, variance is still increasing, yet the modified Levene test returns a p -value of 0.23, indicating no evidence of unequal variance. Since the transformation does not remedy the increasing variance problem and there were no normality problems with the original data, Welch's test for equal means with unequal variance is used on the original measurements. The large p -value of 0.26 from Welch's test indicates there is no reason to believe the three methods of feeding obese cats result in different weight losses. S code to compute Welch's test and its p -value follow.

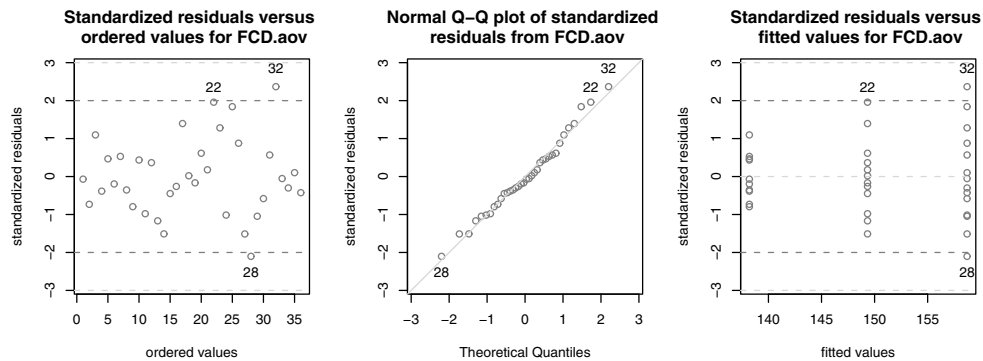


FIGURE 11.12: `checking.plots()` applied to the model `FCD.aov` (`aov(Weight ~ Diet)`) with the `FCD` data frame

```
> attach(FCD)
> ni <- tapply(Weight, Diet, length)
> a <- length(ni)
> si2 <- tapply(Weight, Diet, var)
> wi <- ni/si2
> yb <- tapply(Weight, Diet, mean)
> ytild <- sum(wi*yb)/sum(wi)
> wlamb <- 3*sum((1 - (wi/sum(wi)))^2 / (ni - 1)) / (a^2 - 1)
> dfn <- (a - 1)
> dfd <- 1/wlamb
> W <- sum(wi*(yb - ytild)^2 / (3 - 1)) / (1 + 2/3*(3 - 2)*wlamb)
> W
[1] 1.451544
```

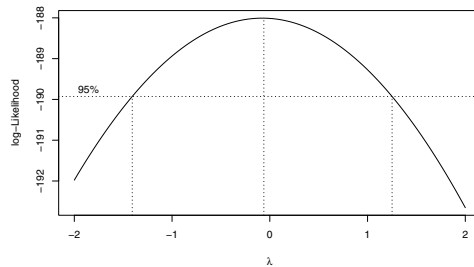



FIGURE 11.13: Box-Cox transformation graph for the model `FCD.aov` (`aov(Weight ~ Diet)`) with the `FCD` data frame

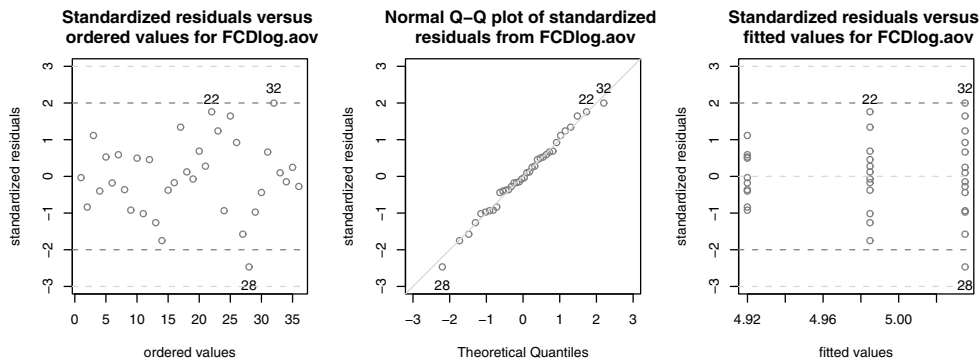


FIGURE 11.14: `checking.plots()` applied to the model `FCDlog.aov` (`aov(log(Weight) ~ Diet)`) using the `FCD` data frame

```
> pvalue <- 1 - pf(W, dfn, dfd)
> pvalue
[1] 0.2561727
> oneway.test(Weight~Diet)
```

One-way analysis of means (not assuming equal variances)

data: Weight and Diet

F = 1.4515, num df = 2.00, denom df = 21.59, p-value = 0.2562

```
> detach(FCD)
```



11.7 Multiple Comparisons of Means

When the null hypothesis for the completely randomized design, $H_0 : \mu_1 = \mu_2 = \dots = \mu_a$, is rejected with an F -test, the test does not indicate which means are different or

how they differ. To do this, several tests are required; however, as noted earlier, repeated application of a test drastically increases type I errors.

Suppose a set of K null hypotheses $H_{0_1}, H_{0_2}, \dots, H_{0_K}$ are to be tested where the overall hypothesis H_0 is true if all of the H_{0_i} s for $i = 1, 2, \dots, K$ are true:

$$H_0 : H_{0_1} \cap H_{0_2} \cap \dots \cap H_{0_K} \tag{11.18}$$

Note that H_0 is rejected if any of the H_{0_i} s is rejected. The **comparison-wise error rate** is the probability of rejecting a particular H_{0_i} in a single test when H_{0_i} is true. Controlling the comparison-wise error rate at the α_c level means that the expected proportion of individual tests that reject H_{0_i} when H_{0_i} is true is α_c . This is the only error rate considered thus far and has previously been denoted as merely α . It is simply the risk one is willing to take of making a type I error in a single test. In contrast to the comparison-wise error rate, the **experiment-wise error rate** is the probability of rejecting at least one of the H_{0_i} s in a series of tests when all of the H_{0_i} s are true, and is denoted α_e . It is the risk of making at least one type I error among the family of comparisons in (11.18). The experiment-wise error rate, α_e , can be evaluated for a family of *independent* tests. Although a set of tests that might be of interest, such as all pairwise differences of a means, are not independent tests, an upper limit on α_e can be established by assuming the tests are independent. There are a total of $m_a = \binom{a}{2} = a(a-1)/2$ tests needed to evaluate all pairwise differences among a means.

The probability of a type I error for any single test is α_c and the probability of a correct decision is $1 - \alpha_c$. If it is assumed that the m_a tests are independent, then the random variable $X =$ number of type I errors has a binomial distribution:

$$X \sim \text{Bin}(n = m_a, \pi = \alpha_c).$$

Since α_e is the probability of making at least one type I error in the family of tests (m_a),

$$\alpha_e = \mathbb{P}(X \geq 1) = 1 - \mathbb{P}(X = 0) = 1 - \binom{m_a}{0} \alpha_c^0 (1 - \alpha_c)^{m_a} = 1 - (1 - \alpha_c)^{m_a}$$

Table 11.5: α_e values for given α_c and various numbers of comparisons

		K =number of independent comparisons				
		2	5	10	20	50
α_c	0.01	0.0199	0.0490	0.0956	0.1821	0.3950
	0.05	0.0975	0.2262	0.4013	0.6415	0.9231
	0.10	0.1900	0.4095	0.6513	0.8784	0.9948

Glancing at Table 11.5, one sees very clearly that for fixed α_c , as K increases, α_e tends to 1. In other words, the probability of making at least one type I error in a series of tests approaches 1 as the number of tests increases. Consequently, multiple comparisons will generally attempt to control α_e , the experiment-wise error rate. To obtain a rough idea of the value of α_e , one can use the Bonferroni inequality $\alpha_e \leq K \cdot \alpha_c$. Likewise, a rough estimate of α_c is α_e/K .

11.7.1 Fisher’s Least Significant Difference

Fisher’s least significant difference (protected LSD) requires an overall F -test of H_0 . If H_0 is rejected, t -tests are used with a common variance estimator (MS_{Error}) for comparisons

of interest. This procedure, despite its appearance, controls neither α_c nor α_e and is not a recommended testing procedure. It is included here for pedagogical reasons only. For pairwise comparisons, group means are considered different if

$$|\bar{Y}_{i\bullet} - \bar{Y}_{j\bullet}| > \underbrace{t_{1-\frac{\alpha_c}{2}; df_{\text{Error}}} \cdot \sqrt{MS_{\text{Error}}}}_{\text{LSD}} \cdot \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \quad (11.19)$$

The $(1 - \alpha_c) \cdot 100\%$ confidence interval on the difference of means based on the LSD is

$$CI_{1-\alpha_c}(\mu_i - \mu_j) = \left[\begin{aligned} &(\bar{Y}_{i\bullet} - \bar{Y}_{j\bullet}) - t_{1-\alpha_c/2; df_{\text{Error}}} \sqrt{MS_{\text{Error}}} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}, \\ &(\bar{Y}_{i\bullet} - \bar{Y}_{j\bullet}) + t_{1-\alpha_c/2; df_{\text{Error}}} \sqrt{MS_{\text{Error}}} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \end{aligned} \right] \quad (11.20)$$

When the number of comparisons is small ($K \leq 5$), the problem of an increasing α_e for using Fisher's LSD can be addressed with the **Bonferroni** method.

The Bonferroni method divides α_c by the total number (K) of comparisons. Means are considered different if the difference of sample means is greater than Bonferroni's significant difference (BSD):

$$|\bar{Y}_{i\bullet} - \bar{Y}_{j\bullet}| > \underbrace{t_{1-\frac{\alpha_c}{2 \cdot K}; df_{\text{Error}}} \cdot \sqrt{MS_{\text{Error}}}}_{\text{BSD}} \cdot \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \quad (11.21)$$

The $(1 - \alpha_e) \cdot 100\%$ confidence interval on the difference of means based on the BSD is

$$CI_{1-\alpha_e}(\mu_i - \mu_j) = \left[\begin{aligned} &(\bar{Y}_{i\bullet} - \bar{Y}_{j\bullet}) - t_{1-\frac{\alpha_c}{2K}; df_{\text{Error}}} \sqrt{MS_{\text{Error}}} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}, \\ &(\bar{Y}_{i\bullet} - \bar{Y}_{j\bullet}) + t_{1-\frac{\alpha_c}{2K}; df_{\text{Error}}} \sqrt{MS_{\text{Error}}} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \end{aligned} \right] \quad (11.22)$$

The experiment-wise error rate using α_c/K can be much less than α_e , thus this method is very conservative and has correspondingly low power.

11.7.2 The Tukey's Honestly Significant Difference

The Tukey's honestly significant difference (HSD) was designed to control α_e . As such, it does a much better job of keeping α_e close to its nominal level than does the Bonferroni procedure. The HSD procedure is based on the studentized range statistic. The studentized range statistic, Q , for a set of treatment means is

$$Q = \frac{\bar{Y}_{\max} - \bar{Y}_{\min}}{\hat{\sigma}/\sqrt{n}} \quad (11.23)$$

The distribution of Q depends on the number of treatments (a) and the degrees of freedom for $\hat{\sigma}$ (MS_{Error}), denoted by ν . In the one-way CRD, $df_{\text{Error}} = N - a$. The notation $q_{1-\alpha; a, \nu}$ denotes the studentized range value with $1 - \alpha$ area to the left with a and ν degrees of freedom, respectively. The S function `qtukey()` returns values from the studentized range distribution. For example, $q_{0.95; 4, 20} = 3.958$ is obtained by entering `qtukey(0.95, 4, 20)`.

The HSD method rejects any pairwise null hypothesis $H_0 : \mu_i = \mu_j$ at the α_e level if

$$|\bar{Y}_{i\bullet} - \bar{Y}_{j\bullet}| > \underbrace{q_{1-\alpha_e; a, \nu} \cdot \frac{\sqrt{MS_{\text{Error}}}}{\sqrt{n}}}_{\text{HSD}} \quad (11.24)$$

Note that

$$\frac{|\bar{Y}_{i\bullet} - \bar{Y}_{j\bullet}|}{\sqrt{MS_{\text{Error}}}\sqrt{\frac{1}{n} + \frac{1}{n}}} = |t| > \frac{q_{1-\alpha_e; a, \nu}}{\sqrt{2}}$$

which implies a confidence interval for $\mu_i - \mu_j$ at the $1 - \alpha_e$ level using the studentized range statistic is written as

$$CI_{1-\alpha_e}(\mu_i - \mu_j) = \left[(\bar{Y}_{i\bullet} - \bar{Y}_{j\bullet}) - \frac{q_{1-\alpha_e; a, \nu}}{\sqrt{2}} \sqrt{MS_{\text{Error}}} \sqrt{\frac{1}{n} + \frac{1}{n}}, \right. \\ \left. (\bar{Y}_{i\bullet} - \bar{Y}_{j\bullet}) + \frac{q_{1-\alpha_e; a, \nu}}{\sqrt{2}} \sqrt{MS_{\text{Error}}} \sqrt{\frac{1}{n} + \frac{1}{n}} \right] \quad (11.25)$$

Strictly speaking, HSD is only applicable to the equal sample size problem. For unequal sample sizes, HSD can be approximated as

$$\text{HSD} \approx \frac{q_{1-\alpha_e; a, \nu}}{\sqrt{2}} \cdot \sqrt{MS_{\text{Error}}} \cdot \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

which is typically conservative compared to the case of equal n_i and n_j .

Note that the critical values used by the LSD, HSD, and BSD procedures for detecting pairwise differences are

$$t_{1-\frac{\alpha}{2}; \nu} \leq \frac{q_{1-\alpha; a, \nu}}{\sqrt{2}} \leq t_{1-\frac{\alpha}{2K}; \nu}$$

This inequality implies that LSD has the most power for detecting differences, followed by HSD and then BSD. Unfortunately, LSD does not control α_e , while HSD and BSD both do. Consequently, of the three methods used to compare pairwise means, HSD is the one recommended because it controls α_e with equal n and is only slightly conservative when n_i and n_j are unequal.

11.7.3 Displaying Pairwise Comparisons

Pairwise comparisons for K means generate $\binom{K}{2} = K(K-1)/2$ tests. A compact method for displaying the results is to

1. Sort the K means in increasing order.
2. Place the labels of those sorted means on a horizontal axis.
3. Draw lines under groups that are not significantly different.

If consecutive groups are not significantly different, use a single line segment under all of such groups. Suppose there are four treatments being studied, which are labeled A, B, C, and D. The diagram



indicates that A and D are not distinguishable from each other, nor are D, C, and B distinguishable from each other. Only A can be distinguished from C and B.

11.8 Other Comparisons among the Means

At times, comparisons other than pairwise are of interest. For example, suppose tires with tread A and tread B are made in South Carolina and tires with tread C and tread D are made in Florida. In this scenario, assuming the stopping distance for a car traveling 60 miles per hour were being measured, one may want to know if there are differences due to tire manufacturing location and would want to test

$$H_0 : \frac{\mu_A + \mu_B}{2} = \frac{\mu_C + \mu_D}{2}.$$

Any linear combination of means $C = \sum_{i=1}^a c_i \mu_i$, where $\sum_{i=1}^a c_i = 0$, is called a **contrast**. An estimate of the contrast $C = \sum_{i=1}^a c_i \mu_i$ can be obtained from the observed data and expressed as $\hat{C} = \sum_{i=1}^a c_i \bar{Y}_{i\bullet}$. A contrast of observed means is an unbiased estimate of the corresponding true treatment means:

$$E\left(\hat{C} = \sum_{i=1}^a c_i \bar{Y}_{i\bullet}\right) = \sum_{i=1}^a c_i \mu_i \quad (11.26)$$

Since the treatment means are independent, the variance of the observed contrast is

$$\text{Var}\left(\hat{C} = \sum_{i=1}^a c_i \bar{Y}_{i\bullet}\right) = \sigma^2 \sum_{i=1}^a \frac{c_i^2}{n_i}. \quad (11.27)$$

Using the standard form of a t -statistic,

$$\frac{\text{unbiased estimator} - \text{hypothesized value}}{\text{standard error of estimator}},$$

a test statistic for testing $H_0 : \sum_{i=1}^a c_i \mu_i = \delta$ is written as

$$t = \frac{\sum_{i=1}^a c_i \bar{Y}_{i\bullet} - \delta}{\sqrt{MS_{\text{Error}}} \cdot \sqrt{\sum_{i=1}^a \frac{c_i^2}{n_i}}} \quad (11.28)$$

which is distributed as a t -distribution with $N - a$ degrees of freedom when H_0 is true. A confidence interval for any contrast is then

$$CI_{1-\alpha} \left(\sum_{i=1}^a c_i \mu_i \right) = \left[\sum_{i=1}^a c_i \bar{Y}_{i\bullet} - t_{1-\frac{\alpha}{2}; N-a} \cdot \sqrt{MS_{\text{Error}}} \cdot \sqrt{\sum_{i=1}^a \frac{c_i^2}{n_i}}, \right. \\ \left. \sum_{i=1}^a c_i \bar{Y}_{i\bullet} + t_{1-\frac{\alpha}{2}; N-a} \cdot \sqrt{MS_{\text{Error}}} \cdot \sqrt{\sum_{i=1}^a \frac{c_i^2}{n_i}} \right] \quad (11.29)$$

The sum of squares can also be computed for a contrast. In particular, the sum of squares for $\sum_{i=1}^a c_i \bar{Y}_{i\bullet}$ is

$$SS_{\hat{C}} = \frac{\left(\sum_{i=1}^a c_i \bar{Y}_{i\bullet} \right)^2}{\sum_{i=1}^a \frac{c_i^2}{n_i}} \quad (11.30)$$

which has 1 degree of freedom. To test if the contrast $C = \sum_{i=1}^a c_i \mu_i$ is zero the ratio $SS_{\hat{C}}/MS_{\text{Error}}$ is formed, which follows an $F_{1, df_{\text{Error}}}$ when H_0 is true.

11.8.1 Orthogonal Contrasts

The contrasts C and D are said to be orthogonal if

$$\sum_{i=1}^a \frac{c_i d_i}{n_i} = 0.$$

Orthogonal contrasts are independent of one another and partition the treatment sum of squares. That is, if one computes the sum of squares for a full set of orthogonal contrasts ($a - 1$ contrasts for a treatments), adding up the $a - 1$ orthogonal contrasts will equal the treatment sum of squares ($SS_{\text{Treatment}}$). Unfortunately, the construction of a complete set of meaningful contrasts is not an easy proposition. Contrasts should be used to answer scientific questions of interest rather than because a complete set of orthogonal contrasts can be computed.

Example 11.4 ▷ *Drosophila* ◁ The data set **Drosophila** contains per diem fecundity (number of eggs laid per female per day for the first 14 days of life) for 25 females from each of three lines of *Drosophila melanogaster*. The three lines are Nonselected (control), Resistant, and Susceptible. The original measurements are from an experiment conducted by R. R. Sokal (Sokal and Rohlf, 1994, p. 237). Test if there are

- (a) Differences between the three genetic lines,
- (b) Differences in fecundity between the Resistant and the Susceptible lines versus the Nonselected line, and
- (c) Fecundity differences between the Resistant and the Susceptible lines.

Solution: The first question (a) seeks to answer if there are differences in the treatment means. In this case, the hypothesis of interest is $H_0 : \mu_{\text{Nonselected}} = \mu_{\text{Resistant}} = \mu_{\text{Susceptible}}$. The second question is typical of experiments with two new treatments and a control. The

null hypothesis for question (b) is equality between the Nonselected line (control) and the Resistant and the Susceptible lines (the two new treatments), written

$$H_0 : \mu_{\text{Nonselected}} = \frac{\mu_{\text{Resistant}} + \mu_{\text{Susceptible}}}{2}.$$

The hypothesis needed to answer question (c) of whether the two treatments (Resistant and Susceptible) are different is written

$$H_0 : \mu_{\text{Resistant}} = \mu_{\text{Susceptible}}.$$

(a) To test $H_0 : \mu_{\text{Nonselected}} = \mu_{\text{Resistant}} = \mu_{\text{Susceptible}}$ versus $H_1 : \mu_i \neq \mu_j$ for some $i \neq j$, an F -test is formed from the ratio of $MS_{\text{Treatment}}/MS_{\text{Error}} = 8.67$ that yields a p -value of 0.0004. Based on the small p -value, the null hypothesis of equal means is rejected. The evidence suggests mean fecundity between lines is different. The values for the ANOVA table needed to test the null hypothesis are provided Table 11.6. S commands to compute the ANOVA table are

```
> attach(Drosophila)
> summary(aov(Fecundity~Line))
              Df Sum Sq Mean Sq F value    Pr(>F)
Line           2 1362.2   681.1   8.6657 0.0004244 ***
Residuals     72 5659.0    78.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 11.6: ANOVA table for model $\text{Fecundity} \sim \text{Line}$ using *Drosophila* data

Source	df	SS	MS	F	p -value
Treatments	$3 - 1 = 2$	1362.2	$\frac{1362.2}{2} = 681.1$	$\frac{681.1}{78.6} = 8.6657$	0.0004
Error	$75 - 3 = 72$	5659.0	$\frac{5659.0}{72} = 78.6$		
Total	74	7021.2			

Before answering (b) and (c), the residuals are examined (not shown) for the model $\text{Fecundity} \sim \text{Line}$ with the function `checking.plots()`. No problems are noted, so the second and third questions can be answered using the orthogonal contrasts

$$C_1 = \mu_{\text{Nonselected}} - \frac{\mu_{\text{Resistant}} + \mu_{\text{Susceptible}}}{2}$$

which has coefficients $c_i = (1, -0.5, -0.5)$, and

$$C_2 = \mu_{\text{Resistant}} - \mu_{\text{Susceptible}}$$

which has coefficients $d_i = (0, 1, -1)$. Contrasts C_1 and C_2 are orthogonal because

$$\sum_{i=1}^a \frac{c_i \cdot d_i}{n_i} = \frac{1 \times 0}{25} + \frac{-0.5 \times 1}{25} + \frac{-0.5 \times -1}{25} = 0.$$

Since there are $a = 3$ treatments, there are two degrees of freedom for a set of orthogonal contrasts. The sum of squares for the first contrast is 1329.0817 and the sum of squares for the second contrast is 33.1298. The sum of squares for treatments is 1362.2115, which equals the sum of the sum of squares for the two orthogonal contrasts: $1329.0817 + 33.1298$. The φ -value for the first contrast (φ -value = 0.0001) provides strong evidence to suggest $\mu_{\text{Nonselected}} \neq \frac{\mu_{\text{Resistant}} + \mu_{\text{Susceptible}}}{2}$. The φ -value for the second contrast (φ -value = 0.518) provides insufficient evidence to reject the null hypothesis $\mu_{\text{Resistant}} = \mu_{\text{Susceptible}}$.

The sums of squares for \widehat{C}_1 and \widehat{C}_2 using (11.30) are computed as

$$SS_{\widehat{C}_1} = \frac{[(1 \times 33.372) + (-0.5 \times 25.256) + (-0.5 \times 23.628)]^2}{\frac{1^2}{25} + \frac{(-.5)^2}{25} + \frac{(-.5)^2}{25}} = 1329.08$$

and

$$SS_{\widehat{C}_2} = \frac{[(0 \times 33.372) + (1 \times 25.256) + (-1 \times 23.628)]^2}{\frac{0^2}{25} + \frac{1^2}{25} + \frac{(-1)^2}{25}} = 33.13$$

Table 11.7: ANOVA table for orthogonal contrasts with *Drosophila*

Source	df	SS	MS	F	φ -value
\widehat{C}_1	1	1329.08	1329.08	16.91	0.0001
\widehat{C}_2	1	33.13	33.13	0.42	0.5182
Treatments	2	1362.21	681.11	8.67	0.0004
Error	72	5659.00	78.60		
Total	74	7021.21			

Note the φ -values in Table 11.7 are individual φ -values. That is, they are not simultaneously correct φ -values. To obtain φ -values adjusted for simultaneous inference or simultaneous confidence intervals, one should use the R package `multcomp`.

The S commands to calculate the values used in the ANOVA table for the contrasts are

```
> MSE <- summary(aov(Fecundity~Line))[[1]][2, 3] #Remove [[1]] for S-PLUS
> SSTreat <- summary(aov(Fecundity~Line))[[1]][1, 2] # [[1]] for R
> ybari <- tapply(Fecundity, Line, mean)
> dfe <- 75 - 3
> ni <- c(25, 25, 25)
> ci <- c(1, -.5, -.5)
> di <- c(0, 1, -1)
> ORTHO <- sum(ci*di/ni) # verify orthogonality
> ORTHO
[1] 0
> SSC1 <- (sum(ci*ybari))^2/sum((ci^2/ni))
> SSC2 <- (sum(di*ybari))^2/sum((di^2/ni))
> OSSC <- c(SSC1, SSC2)
> c(SSC1, SSC2, SSC1+SSC2, SSTreat)
[1] 1329.0817 33.1298 1362.2115 1362.2115
> FC1 <- SSC1/MSE
```



```

> FC2 <- SSC2/MSE
> Fs <-c(FC1, FC2)
> pval <- 1 - pf(Fs, 1, dfe)
> cbind(OSSC, Fs, pval)
      OSSC      Fs      pval
[1,] 1329.0817 16.9099666 0.0001027371
[2,]   33.1298  0.4215120 0.5182493283
> contrasts(Line)[,1] <- ci
> contrasts(Line)[,2] <- di
> CO <- contrasts(Line)
> colnames(CO) <- c("C1", "C2")
> CO
      C1 C2
Nonselected  1.0  0
Resistant   -0.5  1
Susceptible -0.5 -1
> summary(aov(Fecundity~C(Line, CO, 1)+C(Line, CO, 2)))
      Df Sum Sq Mean Sq F value    Pr(>F)
C(Line, CO, 1)  1 1329.1   1329.1  16.9100 0.0001027 ***
C(Line, CO, 2)  1   33.1     33.1  0.4215 0.5182493
Residuals      72 5659.0     78.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

There are several ways to obtain contrasts with `S` by changing the type of contrasts `S` uses. Contrasts settings for `S` include `contr.helmert`, `contr.poly`, `contr.sum`, and `contr.treatment`. R also has the contrast `contr.SAS`, which is not directly available in S-PLUS. The interested reader should refer to the help documentation by typing `?contr.helmert` for more explanation. R uses `contr.treatment` for unordered factors which is not strictly a contrast in its default options. S-PLUS, on the other hand, uses `contr.helmert` as its default contrast. The option `contr.helmert` produces Helmert contrasts, which are orthogonal contrasts when there are an equal number of observations at each of the factor levels. For example, R default contrasts for `Line` are

```

> contrasts(Line) <- contr.treatment(levels(Line)) # R defaults
> contrasts(Line)
      Resistant Susceptible
Nonselected      0          0
Resistant        1          0
Susceptible      0          1

```

To compute Helmert contrasts, type

```

> contrasts(Line) <- contr.helmert(levels(Line))
> contrasts(Line)
      [,1] [,2]
Nonselected  -1  -1
Resistant     1  -1
Susceptible   0   2

```

To compute the sum of squares for the contrasts used in parts (b) and (c), key in

```
> contrasts(Line) <- contr.helmert(levels(Line))[3:1, 2:1]
> contrasts(Line)
      [,1] [,2]
Nonselected    2    0
Resistant     -1    1
Susceptible   -1   -1
```

Note that using coefficients (2, -1, -1) is equivalent to using (1, -0.5, -0.5), since the first set of coefficients is simply a linear combination of the second set of coefficients:

```
> CO <- contrasts(Line)
> colnames(CO) <- c("Contrast 1", "Contrast 2")
> CO
      Contrast 1 Contrast 2
Nonselected      2         0
Resistant       -1         1
Susceptible     -1        -1
> summary(aov(Fecundity~C(Line, CO, 1)+C(Line, CO, 2)))
              Df Sum Sq Mean Sq F value    Pr(>F)
C(Line, CO, 1)  1 1329.1  1329.1 16.9100 0.0001027 ***
C(Line, CO, 2)  1   33.1    33.1  0.4215 0.5182493
Residuals      72 5659.0    78.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To obtain simultaneous p -values and confidence intervals, the following R code should be used:

```
> library(multcomp)
> summary(glht(aov(Fecundity~Line), linfct = mcp(Line = t(CO))))
```

Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: User-defined Contrasts

```
Fit: aov(formula = Fecundity ~ Line)
```

Linear Hypotheses:

	Estimate	Std. Error	t value	p value
Contrast 1 == 0	17.860	4.343	4.112	0.000205 ***
Contrast 2 == 0	1.628	2.508	0.649	0.766699

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(Adjusted p values reported)

```
> CI <- confint(glht(aov(Fecundity~Line), linfct = mcp(Line = t(CO))))
```

```
> CI
```

Simultaneous Confidence Intervals for General Linear Hypotheses

Multiple Comparisons of Means: User-defined Contrasts

```
Fit: aov(formula = Fecundity ~ Line)
```

```
Estimated Quantile = 2.2827
```

```
Linear Hypotheses:
```

	Estimate	lwr	upr
Contrast 1 == 0	17.8600	7.9456	27.7744
Contrast 2 == 0	1.6280	-4.0961	7.3521

```
95% family-wise confidence level
```

Using the R function `barplot2()` from the `gregmisc` package, barplots for the mean of the various lines and contrasts are created with superimposed 95% confidence intervals as well as a graph of the 95% simultaneous confidence intervals from the `multcomp` package with the command `plot(CI)` and shown in Figure 11.15.

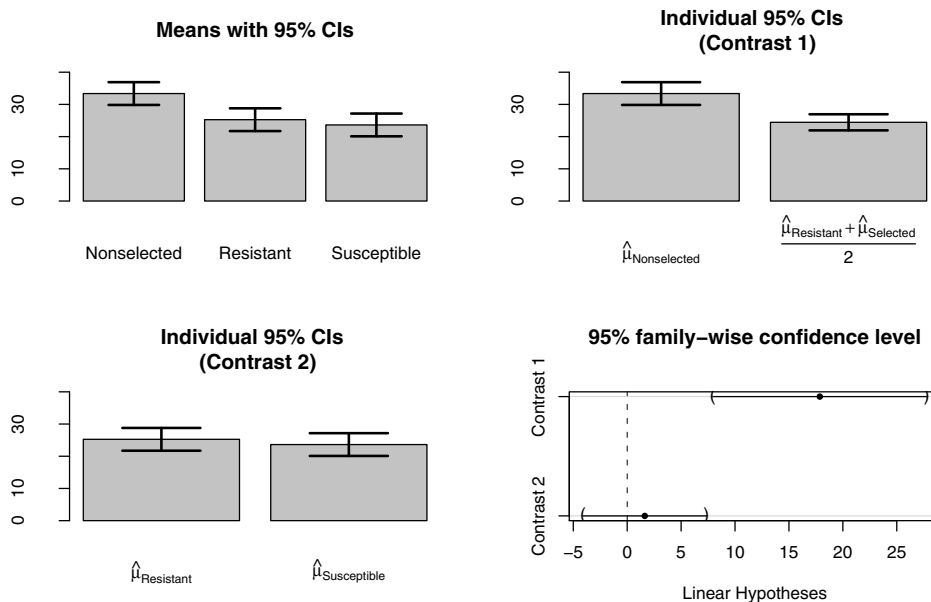


FIGURE 11.15: These graphs give barplots showing the mean fecundity by line, by contrast 1, and by contrast 2 with individual 95% confidence intervals. The bottom right graph displays the simultaneous 95% confidence intervals for contrast 1 and contrast 2.



11.8.2 The Scheffé Method for All Constrasts

The Scheffé method controls the experiment-wise error rate α_e for all possible comparisons, including contrasts, suggested by the data. Consequently, it is the appropriate technique for examining a large number of unplanned comparisons. Its relatively low power limits its legitimate use to data snooping or to investigating contrasts that cannot be handled by other techniques. The Scheffé method is equivalent to the F -test in that the Scheffé method will not find differences in means if the F -test does not reject H_0 . Also, if the F -test does reject H_0 , then there exists at least one comparison that the Scheffé method will declare significant. Unfortunately, finding the comparison(s) that the Scheffé method will declare significant is a process composed entirely of trial and error.

To test the null hypothesis $H_0 : \sum_{i=1}^a c_i \mu_i = 0$ with the Scheffé test statistic S , the ratio $S_{\text{obs}} = \frac{SS_{\hat{C}}/(a-1)}{MS_{\text{Error}}}$ is formed, where $SS_{\hat{C}}$ is as given in (11.30). The null hypothesis is rejected at the α_e level for $S_{\text{obs}} > f_{1-\alpha_e; a-1, \nu}$, where $\nu = df_{\text{Error}}$. For the one-way CRD, $\nu = N - a$. For other models, ν will be different. A confidence interval for an arbitrary contrast, $\sum_{i=1}^a c_i \mu_i$, at the $1 - \alpha_e$ confidence level is

$$CI_{1-\alpha_e} \left(\sum_{i=1}^a c_i \mu_i \right) = \left[\sum_{i=1}^a c_i \bar{Y}_{i\bullet} - \sqrt{(a-1) f_{1-\alpha_e; a-1, \nu}} \cdot \sqrt{MS_{\text{Error}} \sum_{i=1}^a \frac{c_i^2}{n_i}}, \right. \\ \left. \sum_{i=1}^a c_i \bar{Y}_{i\bullet} + \sqrt{(a-1) f_{1-\alpha_e; a-1, \nu}} \cdot \sqrt{MS_{\text{Error}} \sum_{i=1}^a \frac{c_i^2}{n_i}} \right] \quad (11.31)$$

The Scheffé confidence intervals have simultaneous $1 - \alpha_e$ coverage over any set of contrasts.

11.9 Summary of Comparisons of Means

Let the questions to be answered determine the type of contrast that is tested. If the researcher is only interested in determining differences among the means, Tukey’s HSD should be used. Sheffé’s method provides a constant α_e protection for any contrast, which makes it ideal for “data snooping.” Pairwise mean comparisons using any of the methods LSD, BSD, HSD, or Scheffé can be accomplished with the S-PLUS function `multcomp()`. See the S-PLUS help file for specific questions. Tukey’s HSD intervals can be obtained in R using the command `TukeyHSD()`. The R package `MultcompView` renders a graphical representation of the results from `TukeyHSD()`. For simultaneous inference, the R package `multcomp` should be consulted.

Example 11.5 ▷ *Pairwise Mean Comparisons* ◁ Compare all treatment means from Example 11.1 to determine which tire treads have the shortest stopping distance. Use $\alpha = 0.05$ with

- (a) Fisher’s least significant difference
- (b) Bonferroni’s significant difference
- (c) Tukey’s honestly significant difference
- (d) Scheffé’s method

Solution: Each of the methods provides a cutoff value for considering a difference of means significant. The estimated means are $\hat{\mu}_A = 379.67$, $\hat{\mu}_B = 405.16$, $\hat{\mu}_C = 421.67$, and $\hat{\mu}_D = 410.33$. The estimated mean differences with which these values will be compared are

- I. $\hat{\mu}_B - \hat{\mu}_A = \bar{Y}_{2\bullet} - \bar{Y}_{1\bullet} = 25.50$
- II. $\hat{\mu}_C - \hat{\mu}_A = \bar{Y}_{3\bullet} - \bar{Y}_{1\bullet} = 42.00$
- III. $\hat{\mu}_D - \hat{\mu}_A = \bar{Y}_{4\bullet} - \bar{Y}_{1\bullet} = 30.67$
- IV. $\hat{\mu}_C - \hat{\mu}_B = \bar{Y}_{3\bullet} - \bar{Y}_{2\bullet} = 16.50$
- V. $\hat{\mu}_D - \hat{\mu}_B = \bar{Y}_{4\bullet} - \bar{Y}_{2\bullet} = 5.16$
- VI. $\hat{\mu}_D - \hat{\mu}_C = \bar{Y}_{4\bullet} - \bar{Y}_{3\bullet} = -11.33$

(a) Fisher's LSD considers group means significantly different if

$$|\bar{Y}_{i\bullet} - \bar{Y}_{j\bullet}| > \underbrace{t_{1-\frac{\alpha_c}{2}; df_{\text{Error}}} \cdot \sqrt{MS_{\text{Error}}} \cdot \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}_{\text{LSD}}$$

$$\text{LSD} = 2.085 \cdot \sqrt{354.94} \cdot \sqrt{\frac{1}{6} + \frac{1}{6}} = 22.68$$

Comparing this value with the statistics from I–VI indicates that μ_A is significantly different from μ_B , μ_D , and μ_C :

A	B	D	C
	D		

The S-PLUS command to generate Fisher's LSD pairwise confidence intervals is `multicomp(aov(StopDist~tire), method="lsd", error.type="cwe")`.

(b) Bonferroni's significant difference considers group means significantly different if

$$|\bar{Y}_{i\bullet} - \bar{Y}_{j\bullet}| > \underbrace{t_{1-\frac{\alpha_c}{2K}; df_{\text{Error}}} \cdot \sqrt{MS_{\text{Error}}} \cdot \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}_{\text{BSD}}$$

$$\text{BSD} = 2.927 \cdot \sqrt{354.94} \cdot \sqrt{\frac{1}{6} + \frac{1}{6}} = 31.839$$

Comparing this value with the statistics from I–VI indicates that μ_A is significantly different from μ_C . The reader may note that treatment A is not significantly different from treatment B; and, at the same time, treatment B is not significantly different from treatment C. However, treatments A and C are significantly different from each other. When comparing pairwise means, the transitive property does not hold. For orthogonal contrasts, on the other hand, the transitive property will hold.

A	B	D	C
B		D	

The S-PLUS command to generate Bonferroni's significant difference pairwise confidence intervals is `multicomp(aov(StopDist~tire), method="bon")`.

(c) Tukey's honestly significant difference considers group means significantly different if

$$|\bar{Y}_{i\bullet} - \bar{Y}_{j\bullet}| > \underbrace{q_{1-\alpha_e; a, \nu} \cdot \frac{\sqrt{MS_{\text{Error}}}}{\sqrt{n}}}_{\text{HSD}}$$

$$\text{HSD} = 2.799 \cdot \sqrt{354.94} \cdot \sqrt{\frac{1}{6} + \frac{1}{6}} = 30.445$$

Comparing this value with the statistics from I–VI indicates that μ_A is significantly different from μ_D and μ_C :

A	B	D	C

The R code for computing Tukey's HSD's pairwise confidence intervals is

```
> attach(Tire)
> CI <- TukeyHSD(aov(StopDist~tire), which="tire")
> CI
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = StopDist ~ tire)

$tire
      diff      lwr      upr      p adj
B-A 25.500000 -4.9446409 55.94464 0.1213153
C-A 42.000000 11.5553591 72.44464 0.0049515
D-A 30.666667  0.2220258 61.11131 0.0479540
C-B 16.500000 -13.9446409 46.94464 0.4464584
D-B  5.166667 -25.2779742 35.61131 0.9637307
D-C -11.333333 -41.7779742 19.11131 0.7273681

> plot(CI, las=1)
```

Figure 11.16 on the following page shows a graphical representation of the Tukey's HSD confidence intervals calculated above. A slightly different graphical representation of significant differences between group means using Tukey's HSD can be created with the following R code:

```
> library(multcompView)
> multcompBoxplot(StopDist~tire, data=Tire)
```

Figure 11.17 on the next page contains an example of

a matrix of class `multcompTs`, describing the *undifferentiated classes* that identify the other factor levels or items that are not distinct or not significantly different from the *base* of the T ; if two or more levels have the same pattern of significant differences, the two are combined into one T with two *bases*. The resulting T s are similar to the *undifferentiated classes* discussed by Donaghué (2004)

as described in the help file for `multcompTs()` in the R package `multcompView`.

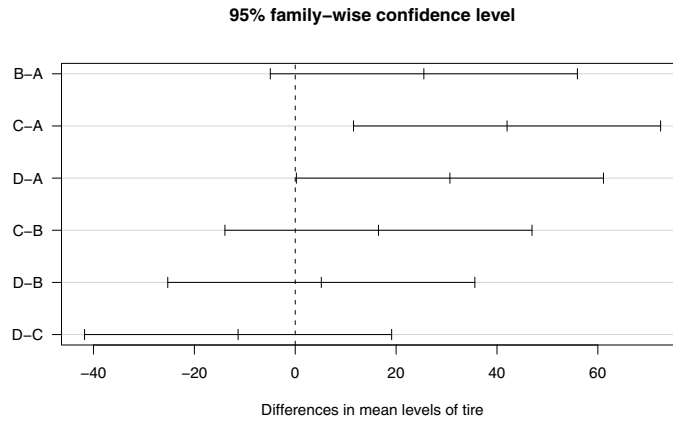


FIGURE 11.16: Graphical representation of confidence intervals based on Tukey's HSD for the model `StopDist ~ tire` using the data frame `Tire`

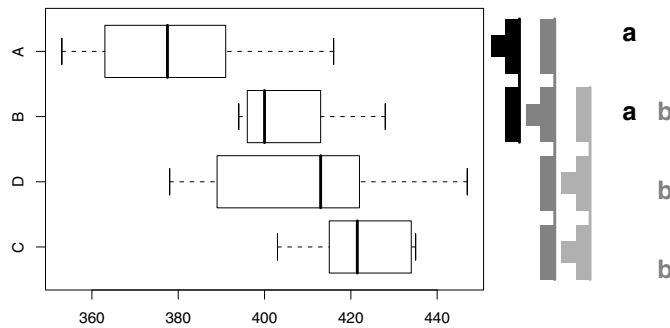


FIGURE 11.17: Multiple comparison boxplot with `multcompTs` differentiating means based on Tukey's HSD for the model `StopDist ~ tire` using the data frame `Tire`

The S-PLUS command to generate Tukey's HSD pairwise confidence intervals is `multcomp(aov(StopDist~tire), method="tukey")`.

(d) If $\sum_{i=1}^a c_i \bar{Y}_{i\bullet}$ is greater than

$$\sqrt{(a-1)f_{1-\alpha_e; a-1, \nu}} \cdot \sqrt{MS_{\text{Error}} \sum_{i=1}^a \frac{c_i^2}{n_i}}$$

Scheffé's method will consider the group means significantly different.

In this case,

$$\sqrt{(a-1)f_{1-\alpha_e; a-1, \nu}} \cdot \sqrt{MS_{\text{Error}} \sum_{i=1}^a \frac{c_i^2}{n_i}} = \sqrt{(4-1) \cdot 3.098} \cdot \sqrt{354.94 \cdot \frac{2}{6}} = 33.162.$$

Comparing this value with the statistics from I–VI indicates that μ_A is significantly different from μ_C :

A	B	D	C
—————			—————

The S-PLUS command to generate Scheffé's significant difference pairwise confidence intervals is `multicomp(aov(StopDist~tire), method="scheffe")`.

S code that can be used to calculate the LSD, BSD, HSD, and Scheffé statistics as well as the pairwise mean differences is

```
> tire.aov <- aov(StopDist~tire)
> alpha.c <- 0.05
> summary(tire.aov)
      Df Sum Sq Mean Sq F value    Pr(>F)
tire      3 5673.1  1891.0  5.3278 0.007316 **
Residuals 20 7098.8   354.9
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> MSE <- summary(aov(tire.aov))[[1]][2, 3]    # For R
> ybari <- tapply(StopDist, tire, mean)
> a <- length(ybari)
> N <- length(StopDist)
> dfe <- N - a
> sort(ybari)
      A      B      D      C
379.6667 405.1667 410.3333 421.6667
> TcritLSD <- qt(1 - alpha.c/2, dfe)
> LSD <- TcritLSD*sqrt(MSE)*sqrt(2/6)
> TcritBON <- qt(1 - alpha.c/(choose(a, 2)*2), dfe)
> BON <- TcritBON*sqrt(MSE)*sqrt(2/6)
> TcritTUK <- qtukey(1 - alpha.c, a, dfe)/sqrt(2)
> HSD <- TcritTUK*sqrt(MSE)*sqrt(2/6)
> CSF <- sqrt((a - 1)*qf(1 - alpha.c, a - 1, dfe))
> SCH <- CSF*sqrt(MSE*2/6)
> c(LSD, BON, HSD, SCH)
[1] 22.68948 31.83892 30.44464 33.16245
> outer(sort(ybari), sort(ybari), "-")
      A      B      D      C
A  0.00000 -25.50000 -30.66667 -42.00000
B 25.50000  0.00000  -5.16667 -16.50000
D 30.66667  5.16667  0.00000 -11.33333
C 42.00000 16.50000 11.33333  0.00000
```

The following R code can be used to duplicate Figure 11.18 on the following page:

```
> library(gregmisc)
> NS <- tapply(StopDist, tire, length)
> SE <- sqrt(MSE)/sqrt(NS)
> t.v <- qt(.975, dfe)
> ci.l <- ybari - t.v*SE
```



```

> ci.u <- ybari + t.v*SE
> barplot2(ybari, plot.ci=TRUE, ci.l=ci.l, ci.u=ci.u, col="skyblue",
+ ylim=c(0, 450), ci.lwd=2)
> title(main="Mean Stopping Distance by Tire \n with Individual 95% CIs")
> detach(Tire)

```

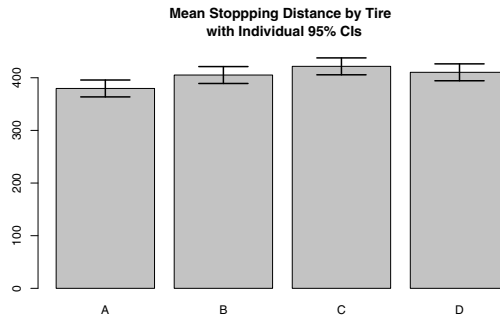


FIGURE 11.18: Barplot of mean stopping distance by tire type with superimposed individual 95% confidence intervals for the `Tire` data frame ■

11.10 Random Effects Model (Variance Components Model)

In the motivational problem Example 11.1, the levels of the factor (tread type) were considered fixed, since only four tread types were available and a decision was sought for the effect of stopping distance for these four types of treads. If, however, the experimenter is interested in a factor that has a large number of possible values and randomly selects a of the possible levels from the population of factor levels, the experiment is modeled as a random effects model. This model is different from the fixed effects model because the levels of the factor are chosen at random. Consequently, inference will apply to the entire population of factor levels, not merely to the a levels in the model. For example, consider a clothing manufacturer that produces work clothes. The strength of the material used in the clothes varies depending on the wool supplier. The manufacturer contracts with a few out of many hundreds of wool suppliers (usually on the basis of price). Since the number of suppliers is very large, by randomly selecting a few suppliers, one can estimate the variability in clothing strength due to suppliers. That is, one is not interested in the particular randomly selected supplier per se. Rather, the goal is to learn something about the suppliers' variability as a whole relative to clothing strength. The statistical model for the one-way random effects remains

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}.$$

In contrast, τ_i is now considered a random variable; whereas in the fixed effect model, it was a parameter. That is, in the random effects model, both τ_i and ε_{ij} are random variables. Consequently the constraint $\sum_{i=1}^a \tau_i = 0$ from the fixed effects model does not apply to the random effects model. The assumptions, then, for the one-way random effects model are

(1) $\varepsilon_{ij} \sim NID(0, \sigma)$.

(2) $\tau_i \sim NID(0, \sigma_\tau)$.

(3) τ_i and ε_{ij} are independent.

Because of assumption number (3), the variance of any observation is $\sigma_{Y_{ij}}^2 = \sigma_{\tau_i}^2 + \sigma^2$. In the random effects model, one is interested in estimating variance components, not in testing treatment means. The reason for this is that the means will vary due to the random nature of selecting the a treatments from the entire population of possible treatments. The partitioning of the sum of squares employed with the fixed effects model is still valid with the random effects model; however, the hypotheses of interest are now

$$H_0 : \sigma_\tau^2 = 0 \quad \text{versus} \quad H_1 : \sigma_\tau^2 > 0$$

which are tested using the ANOVA procedure outlined for the fixed effects model. If the null hypothesis cannot be rejected, $\sigma_\tau^2 = 0$, it is concluded that there are no treatment differences. On the other hand, if the alternative hypothesis is supported, $\sigma_\tau^2 > 0$, the conclusion is that variability exists among treatments.

The test statistic for testing $\sigma_\tau^2 = 0$ is $MS_{\text{Treatment}}/MS_{\text{Error}}$, which follows an $F_{a-1, N-a}$ distribution when the null hypothesis is true. Although the same ANOVA table is used for fixed effects and random effects models, the interpretations are different. The conclusions from a random effects model are not limited to the a treatments used in the computation of the test statistic but rather apply to the entire population of treatments. Estimators for the two variance components when the a treatments have equal sample size n are

$$\hat{\sigma}^2 = MS_{\text{Error}} \quad \text{and} \quad \hat{\sigma}_\tau^2 = \frac{MS_{\text{Treatment}} - MS_{\text{Error}}}{n} \tag{11.32}$$

When treatment sample sizes are unequal, the n in (11.32) is replaced with n' , where

$$n' = \frac{1}{a-1} \sum_{i=1}^a n_i - \frac{\sum_{i=1}^a n_i^2}{\sum_{i=1}^a n_i} \tag{11.33}$$

Example 11.6 ▷ **Frozen Carrots** ◁ A food processing company that uses many hundreds of freezers is studying the variability of its freezers on the texture of frozen carrots. The shear measured in kN on frozen carrots from four randomly selected freezers is shown in Table 11.8 and available in the `food` data frame. The company would like all of its freezers to be homogeneous in order to control the taste of the frozen carrots.

- (a) Test the null hypothesis $H_0 : \sigma_\tau^2 = 0$ for freezers.
- (b) Estimate the component of variance for freezers.

Table 11.8: Shear on frozen carrots by freezer

	1	2	3	4
A	1.96	1.94	1.98	1.92
B	1.82	1.80	1.86	1.84
C	1.92	1.90	1.94	1.90
D	1.90	1.92	1.98	1.96

Solution: The answers are as follows:

(a) The company in this problem is ultimately interested in reducing freezer variability and wants to know if there is more variability in their frozen carrots due to the carrots themselves or due to the numerous freezers used in freezing the carrots.

The hypotheses to be tested are

$$H_0 : \sigma_\tau^2 = 0 \quad \text{versus} \quad H_1 : \sigma_\tau^2 > 0$$

Table 11.9: Frozen carrots ANOVA table

Source of Variation (Source)	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Square (MS)	F	p-value
Treatments	3	0.035675	0.011892	15.681	0.0001878
Error	12	0.009100	0.000758		
Total	15	0.044775			

From Table 11.9, one can see that there is strong evidence to suggest $\sigma_\tau^2 > 0$ (p-value < 0.0002). In other words, there is more variability due to the freezers than variability due to the carrots.

(b) $\hat{\sigma}^2 = MS_{\text{Error}} = 0.000758$

$$\hat{\sigma}_\tau^2 = \frac{MS_{\text{Treatment}} - MS_{\text{Error}}}{n} = \frac{0.011892 - 0.000758}{4} = 0.00278.$$

S Commands:

```
> attach(food)
> summary(aov(shear~freezer))
      Df Sum Sq Mean Sq F value    Pr(>F)
freezer  3 0.035675 0.011892  15.681 0.0001878 ***
Residuals 12 0.009100 0.000758
---
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
> MSC <- summary(aov(shear~freezer))[[1]][1, 3] # omit [[1]] for S-PLUS
> MSE <- summary(aov(shear~freezer))[[1]][2, 3] # omit [[1]] for S-PLUS
> sig2tau <- (MSC - MSE)/4
> sig2tau
[1] 0.002783333
> detach(food)
```



11.11 Randomized Complete Block Design

The t -tests from Sections 9.7.4 and 9.7.6 were used to compare two treatments. Comparisons between the two treatments used the paired t -test when the measurements being compared were related, and consequently more homogeneous. The main idea behind using the paired t -test was to reduce the overall variability of the experiment by pairing observations. When comparing two treatments, whenever the variability within the pairs is smaller than the between pairs variability, detection of the treatment effect is improved by using a paired design. When observations that are homogeneous in some respect are grouped together, the result is referred to as a **block**. Blocks are used in many settings to reduce variability. Some of these include agricultural studies with different strips of land, different litters of animals, and batches of chemical materials. In this section, the paired t -test is generalized to $a \geq 2$ treatments and the resulting design is referred to as a **randomized complete block design**. (Thus, pairing is a special case of blocking where each block is of size two.) The design is called complete because each treatment is used in every block. Instead of treatments being assigned to experimental units, as was the case in the completely randomized design, the randomized complete block design assigns treatments to an equal number of experimental units (usually one) at random within each block. The statistical model used to represent a randomized complete block design (RCBD) is

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij} \text{ for } i = 1, 2, \dots, a \text{ and } j = 1, 2, \dots, b \quad (11.34)$$

where μ is the grand mean; τ_i is the i^{th} treatment effect, which is the difference between the mean response of the i^{th} treatment over all blocks and the grand mean; β_j is the j^{th} block effect, which is the difference between the mean response of the j^{th} block over all treatments and the grand mean; and ε_{ij} are the $NID(0, \sigma)$ error terms. Treatment and block effects are considered fixed effects, defined as deviations from the grand mean so that $\sum_{i=1}^a \tau_i = 0$ and $\sum_{j=1}^b \beta_j = 0$. Note that model (11.34) is a completely additive model, which assumes blocks and treatments do not interact. That is, if treatment one causes the expected response to increase by 3 units ($\tau_1 = 3$), and if the first block decreases the expected response by 1 unit ($\beta_1 = -1$), then the expected response for both treatment and block one is $E(Y_{11}) = \mu + \tau_1 + \beta_1 = \mu + 3 - 1 = \mu + 2$. A RCBD is really a design with two factors, where only one factor (the one measuring the treatment effect) is of interest. The other factor (called a block) is used to reduce the experiment's variability and to enhance its ability to detect treatment differences for the factor of interest. Analysis of the RCBD differs from a two-factor design because the blocking factor is not randomized. This dependence in the blocking factor means there is no theoretical justification for a test of blocks. However, one will often look at the ratio $MS_{\text{Blocks}}/MS_{\text{Error}}$ to get an idea if blocking was beneficial. Just keep in mind that the ratio $MS_{\text{Blocks}}/MS_{\text{Error}}$ does not truly follow an F distribution, as does the ratio $MS_{\text{Treatment}}/MS_{\text{Error}}$. One must remember that blocks should only be used when doing so reduces the overall design variability. To do otherwise reduces the power of the test.

The least squares estimators for the parameters in (11.34) are

$$\hat{\mu} = \bar{Y}_{\bullet\bullet} \quad (11.35)$$

$$\hat{\tau}_i = \bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet} \quad (11.36)$$

$$\hat{\beta}_j = \bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet} \quad (11.37)$$

and the residuals are

$$\hat{\varepsilon}_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \bar{Y}_{i\bullet} - \bar{Y}_{\bullet j} + \bar{Y}_{\bullet\bullet}. \quad (11.38)$$

Each Y_{ij} from (11.34) can be decomposed into four parts by substituting the least squares estimates of μ , τ_i , β_j , and ε_{ij} for the parameters' values:

$$\begin{aligned} Y_{ij} &= \hat{\mu} + \hat{\tau}_i + \hat{\beta}_j + \hat{\varepsilon}_{ij} & (11.39) \\ Y_{ij} &= \bar{Y}_{\bullet\bullet} + (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}) + (\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet}) + (Y_{ij} - \bar{Y}_{i\bullet} - \bar{Y}_{\bullet j} + \bar{Y}_{\bullet\bullet}) \\ (Y_{ij} - \bar{Y}_{\bullet\bullet}) &= (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}) + (\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet}) + (Y_{ij} - \bar{Y}_{i\bullet} - \bar{Y}_{\bullet j} + \bar{Y}_{\bullet\bullet}) \end{aligned}$$

Squaring and summing over $i = 1, 2, \dots, a$ and $j = 1, 2, \dots, b$ gives

$$\sum_{i=1}^a \sum_{j=1}^b (Y_{ij} - \bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^a \sum_{j=1}^b [(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}) + (\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet}) + (Y_{ij} - \bar{Y}_{i\bullet} - \bar{Y}_{\bullet j} + \bar{Y}_{\bullet\bullet})]^2 \quad (11.40)$$

When the right side of (11.40) is expanded, all three cross products sum to zero (which is left to the reader to verify), giving

$$\begin{aligned} \underbrace{\sum_{i=1}^a \sum_{j=1}^b (Y_{ij} - \bar{Y}_{\bullet\bullet})^2}_{SS_{\text{Total}}} &= \underbrace{\sum_{i=1}^a \sum_{j=1}^b (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2}_{SS_{\text{Treatment}}} + \underbrace{\sum_{i=1}^a \sum_{j=1}^b (\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet})^2}_{SS_{\text{Block}}} \\ &\quad + \underbrace{\sum_{i=1}^a \sum_{j=1}^b (Y_{ij} - \bar{Y}_{i\bullet} - \bar{Y}_{\bullet j} + \bar{Y}_{\bullet\bullet})^2}_{SS_{\text{Error}}} \quad (11.41) \end{aligned}$$

The symbolic representation of (11.41) is

$$SS_{\text{Total}} = SS_{\text{Treatment}} + SS_{\text{Block}} + SS_{\text{Error}}.$$

The corresponding degrees of freedom are

$$\underbrace{a \cdot b - 1}_{\text{total } df} = \underbrace{a - 1}_{\text{treatment } df} + \underbrace{b - 1}_{\text{block } df} + \underbrace{(a - 1)(b - 1)}_{\text{error } df}.$$

The mean squares are computed as with the completely randomized design model by dividing each sum of squares by its corresponding degrees of freedom. The expected value of the mean squares, if treatments and blocks are fixed, can be shown to be

$$\begin{aligned} E(MS_{\text{Treatments}}) &= \sigma^2 + \frac{b \cdot \sum_{i=1}^a \tau_i^2}{a - 1} \\ E(MS_{\text{Blocks}}) &= \sigma^2 + \frac{a \cdot \sum_{j=1}^b \beta_j^2}{b - 1} \\ E(MS_{\text{Error}}) &= \sigma^2 \end{aligned}$$

Consequently, to test for no treatment effect, one uses the ratio $MS_{\text{Treatment}}/MS_{\text{Error}}$, which has an F distribution with $(a - 1)$ and $(a - 1)(b - 1)$ degrees of freedom when H_0 is

true. There is no formal test for blocks; however, examining the ratio $MS_{\text{Blocks}}/MS_{\text{Error}}$, and comparing it to an F distribution with $(b - 1)$ and $(a - 1)(b - 1)$ degrees of freedom will give an indication of whether blocking is appropriate. If blocking is not appropriate, then it should be eliminated in future experiments. The ANOVA table for the randomized complete block design is given in Table 11.10.

Table 11.10: ANOVA table for the randomized complete block design

Source of Variation (Source)	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Square (MS)	F
Treatments	$a - 1$	$SS_{\text{Treatment}} = b \cdot \sum_{i=1}^a \hat{\tau}_i^2 \equiv \sum_{i=1}^a \sum_{j=1}^b (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2$	$MS_{\text{Treatment}} = \frac{SS_{\text{Treatment}}}{a - 1}$	$\frac{MS_{\text{Treatment}}}{MS_{\text{Error}}}$
Blocks	$b - 1$	$SS_{\text{Blocks}} = a \cdot \sum_{j=1}^b \hat{\beta}_j^2 \equiv \sum_{i=1}^a \sum_{j=1}^b (\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet})^2$	$MS_{\text{Block}} = \frac{SS_{\text{Block}}}{b - 1}$	
Error	$(a - 1)(b - 1)$	$SS_{\text{Error}} = \sum_{i=1}^a \sum_{j=1}^b \hat{\varepsilon}_{ij}^2 \equiv \sum_{i=1}^a \sum_{j=1}^b (Y_{ij} - \bar{Y}_{i\bullet} - \bar{Y}_{\bullet j} + \bar{Y}_{\bullet\bullet})^2$	$MS_{\text{Error}} = \frac{SS_{\text{Error}}}{(a - 1)(b - 1)}$	
Total	$a \cdot b - 1$	$SS_{\text{Total}} = \sum_{i=1}^a \sum_{j=1}^b (Y_{ij} - \bar{Y}_{\bullet\bullet})^2$		

Possible CRBD Treatment Assignments: Tire Wear Suppose a tire manufacturer is interested in determining tire tread loss after 10,000 miles of driving for the company's best selling four tire models. Four cars and four tires of each tire model are available for the experiment. Let the tire models be denoted with the letters A, B, C, and D and the four cars be denoted as Car1, Car2, Car3, and Car4. One possible design is to assign the four tires of model A, B, C, and D to cars Car1, Car2, Car3, and Car4, respectively. This particular design confounds tire model with car, however. That is, it would not be known whether the differences in tire wear are due to cars or tire model. Another solution might be to use a completely randomized design, but not all tire models will necessarily be used on

all cars. Consider the completely random assignment of tire models to cars given in the S output stored in the variable `tireCRD`. Note that tire model D is never used with Car1, tire model C is never used with Car2, and tire model A is never used with Car3. Further, any variation in model A may simply be due to Car1, Car2, and Car4. Although the completely randomized design averaged out the car effects, it did not eliminate the variance among cars. The randomized complete block design does remove the variability due to cars. One possible assignment of tire models within cars is given under the variable `tireCRBD`:

```
> car <- rep(c("Car1", "Car2", "Car3", "Car4"), c(4, 4, 4, 4))
> tire <- rep(LETTERS[1:4], c(4, 4, 4, 4))
> tireCRD <- sample(tire)
> tireCRBD <- c(sample(LETTERS[1:4]), sample(LETTERS[1:4]),
+ sample(LETTERS[1:4]), sample(LETTERS[1:4]))
> Designs <- cbind(car, tire, tireCRD, tireCRBD)
> Designs
      car  tire tireCRD tireCRBD
[1,] "Car1" "A"  "C"    "B"
[2,] "Car1" "A"  "A"    "D"
[3,] "Car1" "A"  "B"    "C"
[4,] "Car1" "A"  "A"    "A"
[5,] "Car2" "B"  "D"    "A"
[6,] "Car2" "B"  "A"    "B"
[7,] "Car2" "B"  "D"    "C"
[8,] "Car2" "B"  "B"    "D"
[9,] "Car3" "C"  "C"    "A"
[10,] "Car3" "C"  "C"    "B"
[11,] "Car3" "C"  "D"    "D"
[12,] "Car3" "C"  "B"    "C"
[13,] "Car4" "D"  "C"    "B"
[14,] "Car4" "D"  "D"    "D"
[15,] "Car4" "D"  "A"    "A"
[16,] "Car4" "D"  "B"    "C"
```

Example 11.7 ▷ *Tire Wear* ◁ The data frame `TireWear` contains measurements for the amount of tread loss after 10,000 miles of driving in thousandths of an inch. The tread loss from the `TireWear` data frame is presented in tabular form in Table 11.11 on the next page along with the order the tires were assigned to the car in parentheses. Use the values in Table 11.11 to test for treatment (tire model) effects using an additive RCBD.

- Verify that an additive model is appropriate.
- Compute the ANOVA table to test $H_0 : \tau_i = 0$ for all i versus $H_1 : \tau_i \neq 0$ for some i .
- Represent the Y_{ij} values using (11.39).
- Verify graphically that $\varepsilon_{ij} \sim N(0, \sigma)$.
- Determine which tires are different (have the least tread loss) using Tukey's HSD at $\alpha_e = 0.05$.

Solution: The answers are as follows:

- The RCBD is completely additive, and the function `interaction.plot()` is used to verify the reasonableness of the additivity assumption before computing any sums of squares.

Table 11.11: The tread loss from the **TireWear** data frame

	Car1	Car2	Car3	Car4
A	10 (4)	8 (1)	7 (1)	7 (3)
B	9 (1)	8 (2)	7 (2)	5 (1)
C	8 (3)	7 (3)	5 (4)	3 (4)
D	6 (2)	5(4)	3 (3)	3 (2)

Table 11.12: Sums and estimates for Example 11.7

		Blocks				$\bar{Y}_{i\bullet}$	$\hat{\tau}_i = \bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}$
		Car1	Car2	Car3	Car4		
Tire Tread	A	10	8	7	7	8.00	1.6875
	B	9	8	7	5	7.25	0.9375
	C	8	7	5	3	5.75	-0.5625
	D	6	5	3	3	4.25	-2.0625
$\bar{Y}_{\bullet j}$		8.25	7.00	5.50	4.50	$\bar{Y}_{\bullet\bullet} = 6.3125$	
$\hat{\beta}_j = \bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet}$		1.9375	0.6875	-0.8125	-1.8125		

Interaction plots show the relative size of main effects and interaction. The pairs (i, \bar{Y}_{ij}) for all j are plotted, and points in the same block are connected. The roles of blocks and treatments can be reversed, and it is often informative to do so with interaction plots. Parallel lines are indicative of additive designs. Lines that cross should be investigated further. Figure 11.19 on the following page does not suggest any problems with the RCBD's assumption of additivity; however, since graphs are often misleading and their interpretation is subjective, other means of analyzing and evaluating interaction should also be explored.

```
> attach(TireWear)
> par(mfrow=c(1, 2), cex=.8)
> interaction.plot(Treat, Block, Wear, type="b", legend=FALSE)
> interaction.plot(Block, Treat, Wear, type="b", legend=FALSE)
> par(mfrow=c(1, 1), cex=1)
```

The interaction plots suggest both a treatment and a block effect. Another graph that is helpful when there is only one observation per treatment/block combination is the strip plot. Results from using the lattice/Trellis function `stripplot()` are shown in Figure 11.20 on the next page. One can see that tire wear increases with tire models in the order D, C, B, and then A. In a similar fashion, one notes that tire wear in cars increases in the order Car4, Car3, Car2, and then Car1. The graph showing tire wear means due to treatments and blocks using the function `plot.design()` is shown in Figure 11.21 on page 543.

```
> library(lattice) # R
> A <- stripplot(Treat~Wear|Block, layout=c(4, 1))
> B <- stripplot(Block~Wear|Treat, layout=c(4, 1))
> print(A, split=c(1, 1, 1, 2), more=TRUE)
> print(B, split=c(1, 2, 1, 2), more=FALSE)
> plot.design(Wear~Treat+Block)
```

(b) Using the values from Table 11.12 on the previous page, the values for the ANOVA

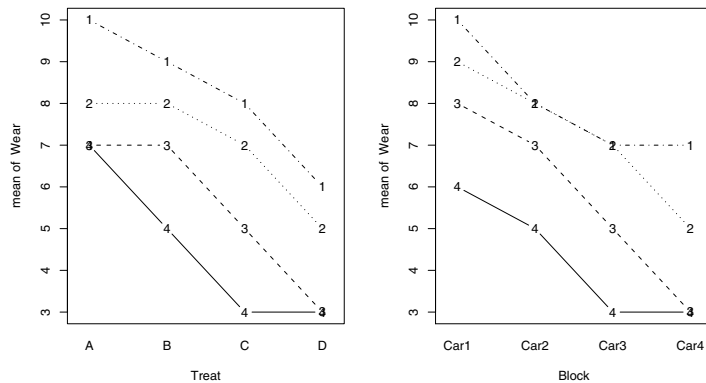


FIGURE 11.19: Left graph shows an interaction plot of blocks and treatments for the response *Wear* where the four blocks, *Car1*, *Car2*, *Car3*, and *Car4*, are denoted with the numbers 1, 2, 3, and 4, respectively, and the treatments shown along the *x*-axis are A, B, C, and D, respectively. The right graph shows an interaction plot of treatments and blocks for the response *Wear* where the four treatments, A, B, C, and D, are denoted with the numbers 1, 2, 3, and 4, respectively, and the blocks shown along the *x*-axis are *Car1*, *Car2*, *Car3*, and *Car4*, respectively.

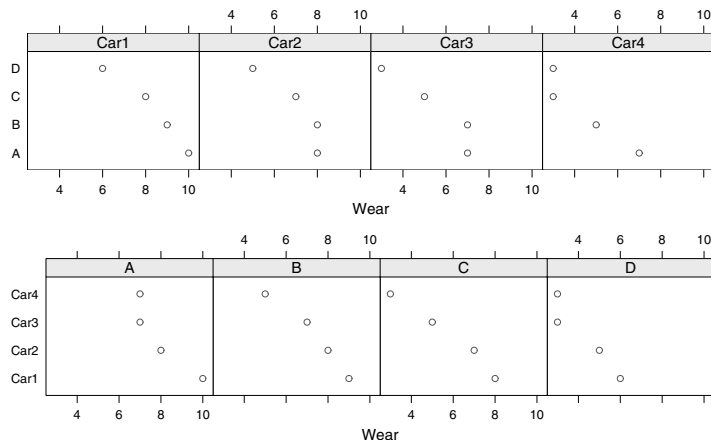


FIGURE 11.20: The graph resulting from the lattice/Trellis function `stripplot()` for Example 11.7

table are

$$\begin{aligned}
 SS_{\text{Treatment}} &= b \cdot \sum_{i=1}^a \hat{\tau}_i^2 = 4(1.6875^2 + 0.9375^2 + (-0.5625)^2 + (-2.0625)^2) \\
 &= 33.1875 \\
 SS_{\text{Block}} &= a \sum_{j=1}^b \hat{\beta}_j^2 = 4(1.9375^2 + 0.6875^2 + (-0.8125)^2 + (-1.8125)^2) \\
 &= 32.6875
 \end{aligned}$$

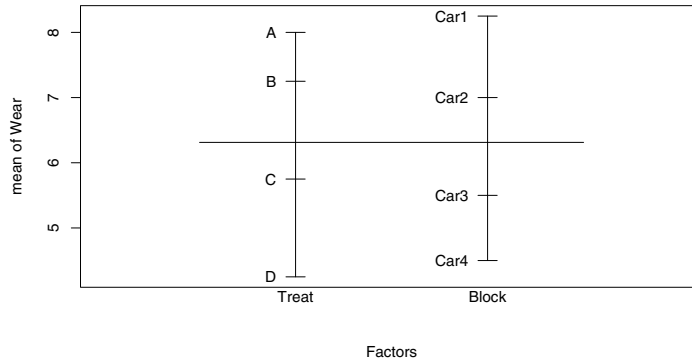


FIGURE 11.21: Tire wear means due to treatments and blocks using the function `plot.design()` for Example 11.7

$$\begin{aligned}
 SS_{\text{Total}} &= \sum_{i=1}^a \sum_{j=1}^b (Y_{ij} - \bar{Y}_{..})^2 \\
 &= (10 - 6.3125)^2 + (9 - 6.3125)^2 + (8 - 6.3125)^2 + \dots + (3 - 6.3125)^2 \\
 &= 69.4375 \\
 SS_{\text{Error}} &= SS_{\text{Total}} - SS_{\text{Treatment}} - SS_{\text{Block}} \\
 &= 69.4375 - 33.1875 - 32.6875 \\
 &= 3.5625
 \end{aligned}$$

Table 11.13: Tire wear ANOVA table

Source	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	ϕ -value
Treatments	3	33.1875	11.062	27.947	0.000068
Blocks	3	32.6875	10.896		
Error	9	3.5625	0.396		
Total	15	69.4375			

To produce the ANOVA table with S, enter

```

> mod.aov <- aov(Wear~Treat+Block)
> summary(mod.aov)
      Df Sum Sq Mean Sq F value    Pr(>F)
Treat   3  33.187  11.062  27.947 6.824e-05 ***
Block   3  32.688  10.896  27.526 7.254e-05 ***
Residuals  9   3.562   0.396
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
    
```

Note that the computer treats the blocking factor as if it were assigned at random and computes a ϕ -value for the blocking factor. The small ϕ -value suggests that blocking is appropriate. Since the ϕ -value = 0.000068, the null hypothesis ($H_0 : \tau_i = 0$) of no treatment effect is rejected.

(c) The Y_{ij} values can be decomposed as follows:

$$\begin{aligned}
 Y_{ij} &= \hat{\mu} + \hat{\tau}_i + \hat{\beta}_j + \hat{\varepsilon}_{ij} \\
 \begin{bmatrix} 10 & 8 & 7 & 7 \\ 9 & 8 & 7 & 5 \\ 8 & 7 & 5 & 3 \\ 6 & 5 & 3 & 3 \end{bmatrix} &= \begin{bmatrix} 6.3125 & 6.3125 & 6.3125 & 6.3125 \\ 6.3125 & 6.3125 & 6.3125 & 6.3125 \\ 6.3125 & 6.3125 & 6.3125 & 6.3125 \\ 6.3125 & 6.3125 & 6.3125 & 6.3125 \end{bmatrix} \\
 &+ \begin{bmatrix} 1.6875 & 1.6875 & 1.6875 & 1.6875 \\ 0.9375 & 0.9375 & 0.9375 & 0.9375 \\ -0.5625 & -0.5625 & -0.5625 & -0.5625 \\ -2.0625 & -2.0625 & -2.0625 & -2.0625 \end{bmatrix} \\
 &+ \begin{bmatrix} 1.9375 & 0.6875 & -0.8125 & -1.8125 \\ 1.9375 & 0.6875 & -0.8125 & -1.8125 \\ 1.9375 & 0.6875 & -0.8125 & -1.8125 \\ 1.9375 & 0.6875 & -0.8125 & -1.8125 \end{bmatrix} \\
 &+ \begin{bmatrix} 0.0625 & -0.6875 & -0.1875 & 0.8125 \\ -0.1875 & 0.0625 & 0.5625 & -0.4375 \\ 0.3125 & 0.5625 & 0.0625 & -0.9375 \\ -0.1875 & 0.0625 & -0.4375 & 0.5625 \end{bmatrix}
 \end{aligned}$$

S code to create the four parts of each Y_{ij} is

```

> yidotbar <- tapply(Wear, Treat, mean)
> ydotjbar <- tapply(Wear, Block, mean)
> gm <- mean(Wear)
> tau_i <- yidotbar - gm
> block_j <- ydotjbar - gm
> GM <- matrix(rep(gm, 16), nrow=4)
> GM
      [,1] [,2] [,3] [,4]
[1,] 6.3125 6.3125 6.3125 6.3125
[2,] 6.3125 6.3125 6.3125 6.3125
[3,] 6.3125 6.3125 6.3125 6.3125
[4,] 6.3125 6.3125 6.3125 6.3125
> treatm <- matrix(rep(tau_i, 4), nrow=4, byrow=FALSE)
> treatm
      [,1] [,2] [,3] [,4]
[1,] 1.6875 1.6875 1.6875 1.6875
[2,] 0.9375 0.9375 0.9375 0.9375
[3,] -0.5625 -0.5625 -0.5625 -0.5625
[4,] -2.0625 -2.0625 -2.0625 -2.0625
> blockm <- matrix(rep(block_j, 4), nrow=4, byrow=TRUE)
> blockm
      [,1] [,2] [,3] [,4]
[1,] 1.9375 0.6875 -0.8125 -1.8125
[2,] 1.9375 0.6875 -0.8125 -1.8125
[3,] 1.9375 0.6875 -0.8125 -1.8125
[4,] 1.9375 0.6875 -0.8125 -1.8125
> residm <- matrix(resid(aov(Wear~Treat+Block)),
+ nrow=4, byrow=FALSE)

```

```
> residm
      [,1] [,2] [,3] [,4]
[1,] 0.0625 -0.6875 -0.1875 0.8125
[2,] -0.1875 0.0625 0.5625 -0.4375
[3,] 0.3125 0.5625 0.0625 -0.9375
[4,] -0.1875 0.0625 -0.4375 0.5625
```

```
> GM+treatm+blockm+residm
      [,1] [,2] [,3] [,4]
[1,] 10 8 7 7
[2,] 9 8 7 5
[3,] 8 7 5 3
[4,] 6 5 3 3
```

The values used in the matrices can also be obtained from using the S function `proj()` (`proj(mod.aov)`).

(d) The residuals from the model `mod.aov` are graphed in Figure 11.22 with the function `checking.plots()` from the PASWR package. The first graph in Figure 11.22 suggests that there is no problem with the independence of errors assumption. The middle graph in Figure 11.22 suggests the errors follow a normal distribution, while the last graph suggests homogeneity of variance is reasonable.

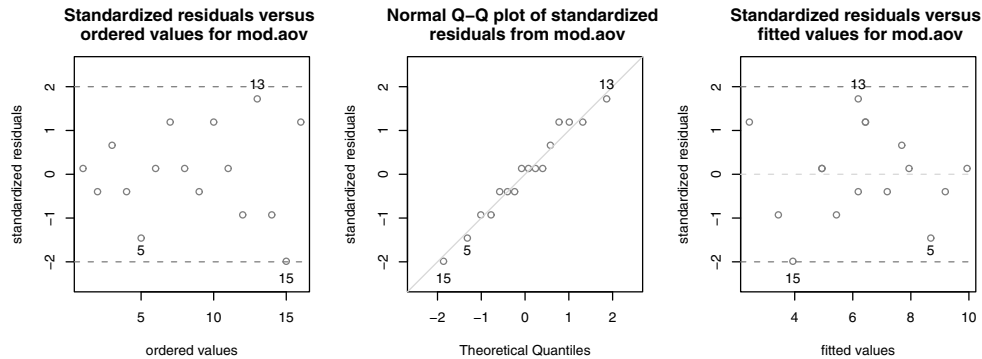


FIGURE 11.22: `checking.plots()` applied to `mod.aov` from Example 11.7

(e) The following R code was used to create simultaneous 95% mean pairwise confidence intervals using Tukey's HSD. The confidence intervals are depicted in Figure 11.23 on the next page.

```
> CI <- TukeyHSD(mod.aov, which="Treat")
> CI
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Wear ~ Treat + Block)
```

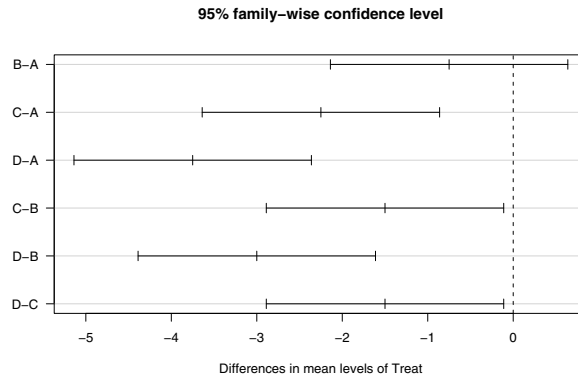
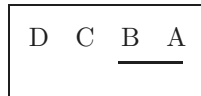


FIGURE 11.23: Simultaneous 95% mean pairwise confidence intervals using Tukey's HSD from Example 11.7

```
$Treat
      diff      lwr      upr    p adj
B-A -0.75 -2.138820  0.6388204 0.3838264
C-A -2.25 -3.638820 -0.8611796 0.0031175
D-A -3.75 -5.138820 -2.3611796 0.0000699
C-B -1.50 -2.888820 -0.1111796 0.0343452
D-B -3.00 -4.388820 -1.6111796 0.0003981
D-C -1.50 -2.888820 -0.1111796 0.0343452
```

```
> plot(CI, las=1)
> detach(TireWear)
```



Tire D is significantly better (less wear) than tires C, B, and A. Tire C is significantly better than tires B and A, and tires B and A are not significantly different from one another. Figure 11.24 shows a barplot of the mean wear by tire with superimposed individual 95% confidence intervals.

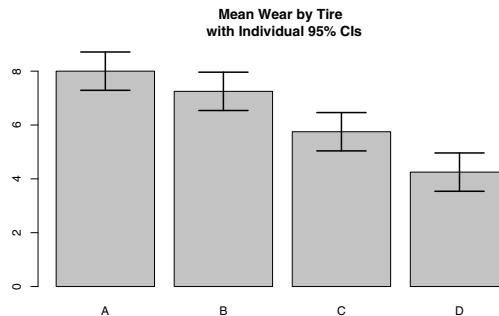


FIGURE 11.24: Barplot of the mean wear by tire with superimposed individual 95% confidence intervals from Example 11.7

11.12 Two-Factor Factorial Design

The one-way analysis of variance evaluated a single factor that had a levels. When a study involves more than one factor, say two fixed factors A and B , with a and b levels, respectively, there are a total of $a \times b = ab$ treatment combinations that need to be analyzed. An efficient method to analyze the ab treatments is with a factorial design. Such a design supplies information about all of the factors in a more efficient fashion than one factor at a time experiments and avoids potentially misleading conclusions that are possible from single-factor designs when interactions are present.

A factorial design with two fixed factors A and B , each with a and b levels, respectively, will typically have n experimental units for each of the ab treatment combinations. The ab treatments are randomly assigned to the $N = abn$ experimental units resulting in a completely randomized design. A general layout for observations from a two-factor factorial design is presented in Table 11.14.

Table 11.14: Layout for observations in a two-factor factorial design

		Factor B				
		1	2	...	b	
Factor A	1	$Y_{111}, Y_{112}, \dots, Y_{11n}$	$Y_{121}, Y_{122}, \dots, Y_{12n}$...	$Y_{1b1}, Y_{1b2}, \dots, Y_{1bn}$	$\bar{Y}_{1..}$
	2	$Y_{211}, Y_{212}, \dots, Y_{21n}$	$Y_{221}, Y_{222}, \dots, Y_{22n}$...	$Y_{2b1}, Y_{2b2}, \dots, Y_{2bn}$	$\bar{Y}_{2..}$
	\vdots	\vdots	\vdots		\vdots	\vdots
	a	$Y_{a11}, Y_{a12}, \dots, Y_{a1n}$	$Y_{a21}, Y_{a22}, \dots, Y_{a2n}$...	$Y_{ab1}, Y_{ab2}, \dots, Y_{abn}$	$\bar{Y}_{a..}$
		$\bar{Y}_{.1.}$	$\bar{Y}_{.2.}$...	$\bar{Y}_{.b.}$	$\bar{Y}_{...}$

The observations from a two-factor factorial design are described by the linear model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk} \quad \text{for } i = 1, \dots, a; j = 1, \dots, b; k = 1, \dots, n \quad (11.42)$$

where μ is the overall mean effect, α_i is the effect of the i^{th} row factor A , β_j is the effect of the j^{th} column factor B , $\alpha\beta_{ij}$ is the effect of the interaction between α_i and β_j , and ε_{ijk} is a random error. Note that $\alpha\beta$ is not $\alpha \cdot \beta$ but rather a single term. Both α_i and β_j are assumed to be fixed with the constraints

$$\sum_{i=1}^a \alpha_i = 0; \quad \sum_{j=1}^b \beta_j = 0; \quad \sum_{i=1}^a \alpha\beta_{ij} = \sum_{j=1}^b \alpha\beta_{ij} = 0. \quad (11.43)$$

That is, the treatment effects are defined as deviations from the overall mean. Given these assumptions, the least squares estimators for the parameters in the two-factor factorial design are

$$\begin{aligned} \hat{\alpha}_i &= \bar{Y}_{i..} - \bar{Y}_{...} & \hat{\beta}_j &= \bar{Y}_{.j.} - \bar{Y}_{...} \\ \hat{\alpha\beta}_{ij} &= \bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...} & \hat{\varepsilon}_{ijk} &= Y_{ijk} - \bar{Y}_{ij.} \end{aligned}$$

Sums of Squares Each Y_{ijk} from (11.42) can be decomposed into five parts by substituting the least squares estimates of μ , α_i , β_j , $\alpha\beta_{ij}$, and ε_{ijk} for the parameters' values:

$$Y_{ijk} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\alpha\beta}_{ij} + \hat{\varepsilon}_{ijk}$$

$$Y_{ijk} = \bar{Y}_{\dots} + (\bar{Y}_{i\bullet\bullet} - \bar{Y}_{\dots}) + (\bar{Y}_{\bullet j\bullet} - \bar{Y}_{\dots}) + (\bar{Y}_{ij\bullet} - \bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet j\bullet} + \bar{Y}_{\dots}) + (Y_{ijk} - \bar{Y}_{ij\bullet})$$

which implies that

$$(Y_{ijk} - \bar{Y}_{\dots}) = (\bar{Y}_{i\bullet\bullet} - \bar{Y}_{\dots}) + (\bar{Y}_{\bullet j\bullet} - \bar{Y}_{\dots}) + (\bar{Y}_{ij\bullet} - \bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet j\bullet} + \bar{Y}_{\dots}) + (Y_{ijk} - \bar{Y}_{ij\bullet}). \quad (11.44)$$

Squaring (11.44) and summing over $i = 1, \dots, a$; $j = 1, \dots, b$; and $k = 1, \dots, n$ gives

$$\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (Y_{ijk} - \bar{Y}_{\dots})^2 = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n [(\bar{Y}_{i\bullet\bullet} - \bar{Y}_{\dots}) + (\bar{Y}_{\bullet j\bullet} - \bar{Y}_{\dots}) + (\bar{Y}_{ij\bullet} - \bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet j\bullet} + \bar{Y}_{\dots}) + (Y_{ijk} - \bar{Y}_{ij\bullet})]^2 \quad (11.45)$$

When the right side of (11.45) is expanded, all four cross products sum to zero (which is left to the reader to verify), giving

$$\underbrace{\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (Y_{ijk} - \bar{Y}_{\dots})^2}_{SS_{\text{Total}}} = \underbrace{bn \sum_{i=1}^a (\bar{Y}_{i\bullet\bullet} - \bar{Y}_{\dots})^2}_{SS_A} + \underbrace{an \sum_{j=1}^b (\bar{Y}_{\bullet j\bullet} - \bar{Y}_{\dots})^2}_{SS_B}$$

$$+ \underbrace{n \sum_{i=1}^a \sum_{j=1}^b (\bar{Y}_{ij\bullet} - \bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet j\bullet} + \bar{Y}_{\dots})^2}_{SS_{AB}} + \underbrace{\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (Y_{ijk} - \bar{Y}_{ij\bullet})^2}_{SS_{\text{Error}}} \quad (11.46)$$

That is,

$$SS_{\text{Total}} = SS_A + SS_B + SS_{AB} + SS_{\text{Error}} \quad (11.47)$$

The corresponding degrees of freedom are

$$\underbrace{abn - 1}_{\text{total df}} = \underbrace{a - 1}_{A \text{ df}} + \underbrace{b - 1}_{B \text{ df}} + \underbrace{(a - 1)(b - 1)}_{AB \text{ interaction df}} + \underbrace{ab(n - 1)}_{\text{Error df}} \quad (11.48)$$

The mean squares are computed by dividing each sum of squares by its degrees of freedom. The expected value of the mean squares, with fixed factors A and B , can be shown to be

$$E(MS_A) = \sigma^2 + \frac{bn \sum_{i=1}^a \alpha_i^2}{a - 1} \quad E(MS_B) = \sigma^2 + \frac{an \sum_{j=1}^b \beta_j^2}{b - 1}$$

$$E(MS_{AB}) = \sigma^2 + \frac{n \sum_{i=1}^a \sum_{j=1}^b \alpha\beta_{ij}^2}{(a - 1)(b - 1)} \quad E(MS_{\text{Error}}) = \sigma^2$$

Consequently, to test for A and B main effects as well as the interaction between A and B , the corresponding mean square is divided by the MS_{Error} . The ANOVA table for a two-factor design is given in Table 11.15 on the facing page. The formal hypotheses for testing for factor A treatment effects, factor B treatment effects, and the interaction between factor A and factor B are written, respectively, as

Factor A	Factor B	Interaction
$H_0 : \alpha_i = 0 \quad \text{for all } i$	$H_0 : \beta_j = 0 \quad \text{for all } j$	$H_0 : \alpha\beta_{ij} = 0 \quad \text{for all } (i, j)$
$H_1 : \alpha_i \neq 0 \quad \text{for some } i$	$H_1 : \beta_j \neq 0 \quad \text{for some } j$	$H_1 : \alpha\beta_{ij} \neq 0 \quad \text{for some } (i, j)$

Table 11.15: ANOVA table for two-factor factorial design

Source	df	SS	MS	F
A	$a - 1$	$SS_A = bn \sum_{i=1}^a (\bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet\bullet\bullet})^2$	$MS_A = \frac{SS_A}{a-1}$	$\frac{MS_A}{MS_{\text{Error}}}$
B	$b - 1$	$SS_B = an \sum_{j=1}^b (\bar{Y}_{\bullet j\bullet} - \bar{Y}_{\bullet\bullet\bullet})^2$	$MS_B = \frac{SS_B}{b-1}$	$\frac{MS_B}{MS_{\text{Error}}}$
AB	$(a - 1)(b - 1)$	$SS_{AB} = n \sum_{i=1}^a \sum_{j=1}^b (\bar{Y}_{ij\bullet} - \bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet j\bullet} + \bar{Y}_{\bullet\bullet\bullet})^2$	$MS_{AB} = \frac{SS_{AB}}{(a-1)(b-1)}$	$\frac{MS_{AB}}{MS_{\text{Error}}}$
Error	$ab(n - 1)$	$SS_{\text{Error}} = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (Y_{ijk} - \bar{Y}_{ij\bullet})^2$	$MS_{\text{Error}} = \frac{SS_{\text{Error}}}{ab(n-1)}$	
Total	$abn - 1$	$SS_{\text{Total}} = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (Y_{ijk} - \bar{Y}_{\bullet\bullet\bullet})^2$		

Example 11.8 ▷ *Television Tube Screen Brightness* ◁ The data in Table 11.16 are taken from Hicks (1956) where an experiment was designed to study the effect of glass type and phosphor type on the brightness of a television tube screen. The measured variable was the current in microamperes (μA) necessary to produce a certain level of brightness. The higher the μA required to produce a given brightness, the poorer are the tube screen characteristics. That is, optimal characteristics are obtained when the response (μA) is small. Analyze the data using a two-factor factorial design.

Table 11.16: Data from Hicks (1956) used in Example 11.8

		Phosphor		
		A	B	C
Glass	I	280, 290, 285	300, 310, 295	270, 285, 290
	II	230, 235, 240	260, 240, 235	220, 225, 230

- (a) Read the data into S.
- (b) Graphically examine the data.
- (c) Fill in the missing values to complete Table 11.17 on the following page.
- (d) Create a two-way ANOVA table using the information from Table 11.17 on the next page and verify your answers using the function `anova()`.
- (e) Analyze the residuals and comment on whether the model from (11.42) fits the data.
- (f) Is there significant interaction between glass type and phosphor type?

Table 11.17: Two-factor factorial design table to complete for (c) of Example 11.8

	Phosphor A		Phosphor B		Phosphor C		$\hat{\alpha}_i =$ $\bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet\bullet\bullet}$
Glass I	$\bar{Y}_{11\bullet} =$		$\bar{Y}_{12\bullet} =$		$\bar{Y}_{13\bullet} =$		$\hat{\alpha}_1 =$
	$\hat{\alpha}_{11} =$		$\hat{\alpha}_{12} =$		$\hat{\alpha}_{13} =$		
Glass II	$\bar{Y}_{21\bullet} =$		$\bar{Y}_{22\bullet} =$		$\bar{Y}_{23\bullet} =$		$\hat{\alpha}_2 =$
	$\hat{\alpha}_{21} =$		$\hat{\alpha}_{22} =$		$\hat{\alpha}_{23} =$		
$\bar{Y}_{\bullet j\bullet}$	$\bar{Y}_{\bullet 1\bullet} =$		$\bar{Y}_{\bullet 2\bullet} =$		$\bar{Y}_{\bullet 3\bullet} =$		
$\hat{\beta}_j = \bar{Y}_{\bullet j\bullet} - \bar{Y}_{\bullet\bullet\bullet}$	$\hat{\beta}_1 =$		$\hat{\beta}_2 =$		$\hat{\beta}_3 =$		$\bar{Y}_{\bullet\bullet\bullet} =$

- (g) Using $\alpha_e = 0.05$, compute Tukey's HSD 95% confidence intervals to determine which combination of glass type and phosphor type require the least μA . Create a graph of the resulting confidence intervals as well as a barplot of the individual means for the six treatment combinations with superimposed 95% individual confidence intervals.

Solution: The answers are as follows:

- (a) The numbers from Table 11.16 on the preceding page are read into the variable `Microamps` and the factors `Glass` and `Phosphor` are created as follows:

```
> Microamps <- c(280, 290, 285, 300, 310, 295, 270, 285, 290, 230,
+ 235, 240, 260, 240, 235, 220, 225, 230)
> Glass <- factor(c(rep("Glass I", 9), rep("Glass II", 9)))
> Phosphor <- factor(rep(rep(c(rep("Phosphor A", 3),
+ rep("Phosphor B", 3), rep("Phosphor C", 3)), 2)))
```

- (b) The function `twoway.plots(Microamps, Glass, Phosphor)` is used to examine the data, and the results are shown in Figure 11.25 on the next page. From Figure 11.25, glass type appears important, and the lines in the interaction plot are nearly parallel, suggesting interaction between the two factors is not significant.

- (c) The values to fill in Table 11.17 are computed using the S function `model.tables()` as follows:

```
> mod.TVB <- aov(Microamps ~ Glass + Phosphor + Glass:Phosphor)
> model.tables(mod.TVB, type="means")
Tables of means
Grand mean
```

262.2222

```

Glass
Glass I Glass II
289.44 235.00

Phosphor
Phosphor A Phosphor B Phosphor C
260.00 273.33 253.33

Glass:Phosphor
  Phosphor
Glass   Phosphor A Phosphor B Phosphor C
Glass I 285.00 301.67 281.67
Glass II 235.00 245.00 225.00
> model.tables(mod.TVB, type="effects")
Tables of effects
    
```

```

Glass
Glass I Glass II
27.222 -27.222

Phosphor
Phosphor A Phosphor B Phosphor C
-2.222 11.111 -8.889
    
```

```

Glass:Phosphor
  Phosphor
Glass   Phosphor A Phosphor B Phosphor C
Glass I -2.2222 1.1111 1.1111
Glass II 2.2222 -1.1111 -1.1111
    
```

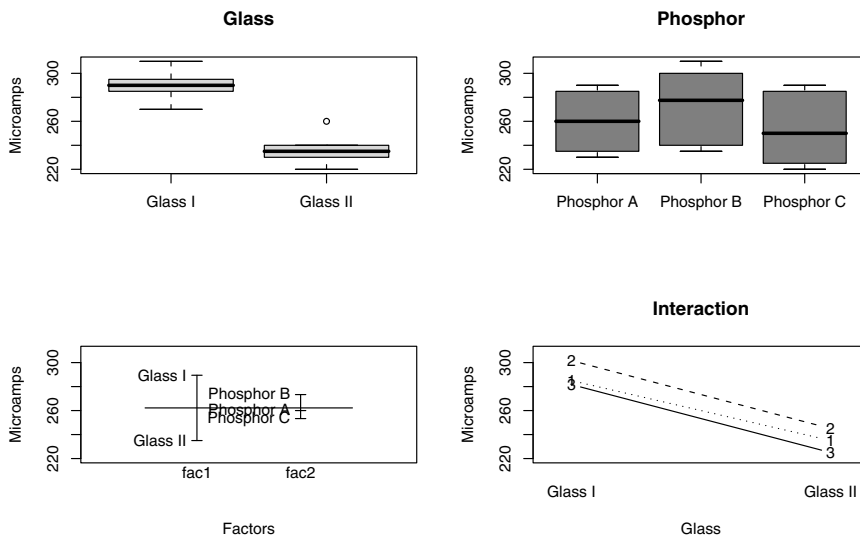


FIGURE 11.25: Graphs resulting from using `twoway.plots(Microamps, Glass, Phosphor)` for Example 11.8

Table 11.18: Two-factor factorial design table COMPLETED for (c) of Example 11.8

	Phosphor A	Phosphor B	Phosphor C	$\bar{Y}_{i\bullet\bullet}$	$\hat{\alpha}_i = \bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet\bullet\bullet}$
Glass I	$\bar{Y}_{11\bullet} = 285$ $\hat{\alpha}\hat{\beta}_{11} = -2.2222$	$\bar{Y}_{12\bullet} = 301.67$ $\hat{\alpha}\hat{\beta}_{12} = 1.1111$	$\bar{Y}_{13\bullet} = 281.67$ $\hat{\alpha}\hat{\beta}_{13} = 1.1111$	$\bar{Y}_{1\bullet\bullet} = 289.44$	$\hat{\alpha}_1 = 27.222$
Glass II	$\bar{Y}_{21\bullet} = 235$ $\hat{\alpha}\hat{\beta}_{21} = 2.2222$	$\bar{Y}_{22\bullet} = 245$ $\hat{\alpha}\hat{\beta}_{22} = -1.1111$	$\bar{Y}_{23\bullet} = 225$ $\hat{\alpha}\hat{\beta}_{23} = -1.1111$	$\bar{Y}_{2\bullet\bullet} = 235$	$\hat{\alpha}_2 = -27.222$
$\bar{Y}_{\bullet j\bullet}$	$\bar{Y}_{\bullet 1\bullet} = 260$	$\bar{Y}_{\bullet 2\bullet} = 273.33$	$\bar{Y}_{\bullet 3\bullet} = 253.33$		
$\hat{\beta}_j = \bar{Y}_{\bullet j\bullet} - \bar{Y}_{\bullet\bullet\bullet}$	$\hat{\beta}_1 = -2.2222$	$\hat{\beta}_2 = 11.1111$	$\hat{\beta}_3 = -8.889$		$\bar{Y}_{\bullet\bullet\bullet} = 262.22$

(d) Using the results from (c), the sums of squares for the ANOVA table are computed and displayed in Table 11.19 on the next page.

$$SS_A = bn \sum_{i=1}^a (\bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet\bullet\bullet})^2 = bn \sum_{i=1}^a \hat{\alpha}_i^2$$

$$= 3 \cdot 3 \cdot [27.222^2 + (-27.222)^2] = 13338.9$$

$$SS_B = an \sum_{j=1}^b (\bar{Y}_{\bullet j\bullet} - \bar{Y}_{\bullet\bullet\bullet})^2 = an \sum_{j=1}^b \hat{\beta}_j^2$$

$$= 2 \cdot 3 \cdot [(-2.2222)^2 + (11.1111)^2 + (-8.889)^2] = 1244.4$$

$$SS_{AB} = n \sum_{i=1}^a \sum_{j=1}^b (\bar{Y}_{ij\bullet} - \bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet j\bullet} + \bar{Y}_{\bullet\bullet\bullet})^2 = n \sum_{i=1}^a \sum_{j=1}^b \hat{\alpha}\hat{\beta}_{ij}^2$$

$$= 3 \cdot [(-2.2222)^2 + (1.1111)^2 + \dots + (-1.1111)^2] = 44.4$$

$$SS_{\text{Error}} = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (Y_{ijk} - \bar{Y}_{ij\bullet})^2$$

$$= [(280 - 285)^2 + (290 - 285)^2 + \dots + (230 - 225)^2] = 833.3$$

$$SS_{\text{Total}} = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (Y_{ijk} - \bar{Y}_{\bullet\bullet\bullet})^2$$

$$= [(280 - 262.2222)^2 + (290 - 262.2222)^2 + \dots + (230 - 262.2222)^2] = 15461.11$$

The values for Table 11.19 on the facing page are verified with the S function `anova()`:

Table 11.19: ANOVA table for two-factor factorial design for Example 11.8

Source	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>
Glass	1	13338.9	13338.9	192.08
Phosphor	2	1244.4	622.2	8.96
Glass:Phosphor	2	44.4	22.2	0.32
Residuals	12	833.3	69.4	
Total	17	15461.0		

```
> anova(mod.TVB)
Analysis of Variance Table
```

Response: Microamps

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Glass	1	13338.9	13338.9	192.08	9.568e-09	***
Phosphor	2	1244.4	622.2	8.96	0.004162	**
Glass:Phosphor	2	44.4	22.2	0.32	0.732158	
Residuals	12	833.3	69.4			

 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(e) The residuals from fitting the data to model (11.42) are analyzed using the function `checking.plots()` and shown in Figure 11.26. The first graph in Figure 11.26 is not relevant because no time component is present in the data, the second graph suggests normality is reasonable and the third graph indicates homogeneity of variance is plausible. Consequently, a two-factor factorial model seems to be a reasonable model for the data on hand.

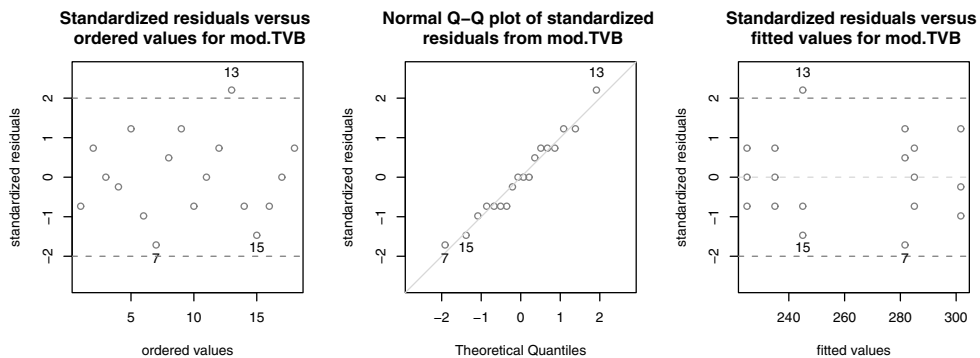


FIGURE 11.26: Graphs resulting from using `checking.plots()` on the model `mod.TVB` from Example 11.8

(f) To assess possible interaction between the factors glass and phosphor, two interaction plots of the same data are created with the following R code and shown in Figure 11.27 on the following page. Since the lines in Figure 11.27 are roughly parallel in both plots, it is

reasonable to assume the two factors, glass and phosphor, do not interact.

```
> par(mfrow=c(1, 2), pty="s")
> interaction.plot(Glass, Phosphor, Microamps, type="b", legend=FALSE)
> interaction.plot(Phosphor, Glass, Microamps, type="b", legend=FALSE)
> par(mfrow=c(1, 1), pty="m")
```

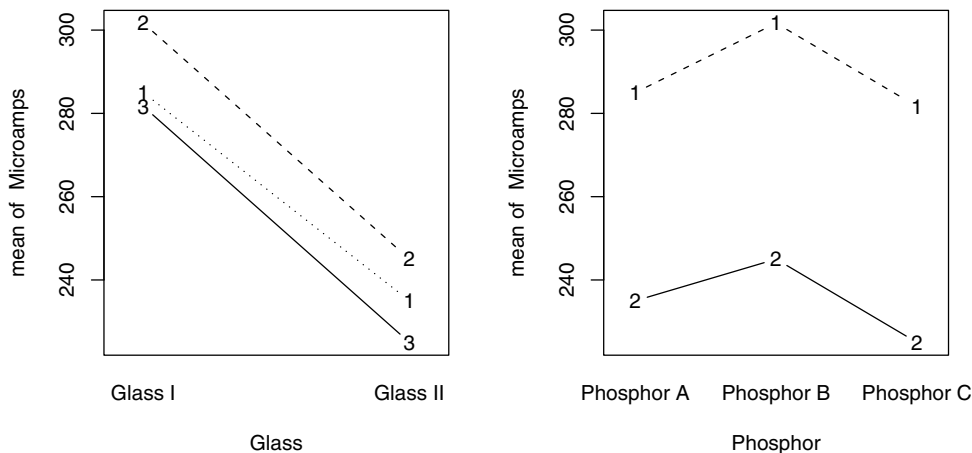


FIGURE 11.27: Left graph shows an interaction plot of glass and phosphor for the response μA where the three types of phosphor, Phosphor A, Phosphor B, and Phosphor C, are denoted with the numbers 1, 2, and 3, respectively, and the glasses depicted along the x -axis are Glass I and Glass II, respectively. The right graph shows an interaction plot of phosphor and glass for the response μA where the two types of glass, Glass I, and Glass II, are denoted with the numbers 1 and 2, respectively, and the phosphors depicted along the x -axis are Phosphor A, Phosphor B, and Phosphor C, respectively.

(g) Tukey 95% confidence intervals are computed for the $\binom{6}{2} = 15$ pairwise differences in mean treatment combinations of glass type and phosphor type using the R function `TukeyHSD()`. The code and output follow with a graph of the confidence intervals shown in Figure 11.28 on the next page. Based on the Tukey HSD confidence intervals, Glass II with either Phosphor A or Phosphor C should be used in the construction of television picture tubes since these combinations require the least μA . Note that the mean for Glass II with Phosphor C is less than the mean for Glass II with Phosphor A but not statistically different.

```
> par(mar=c(5.1, 10.1, 4.1, 2.1), cex.axis=.5)
> CI <- TukeyHSD(mod.TVB)
> plot(CI, las=1)
> par(mar=c(5.1, 4.1, 4.1, 2.1), cex.axis=1)
```

The function `barplot2()` from the `gregmisc` package is used to create a barplot showing the six treatment combination means with individual superimposed 95% confidence intervals using the following R code, with the results displayed in Figure 11.29 on the facing page:

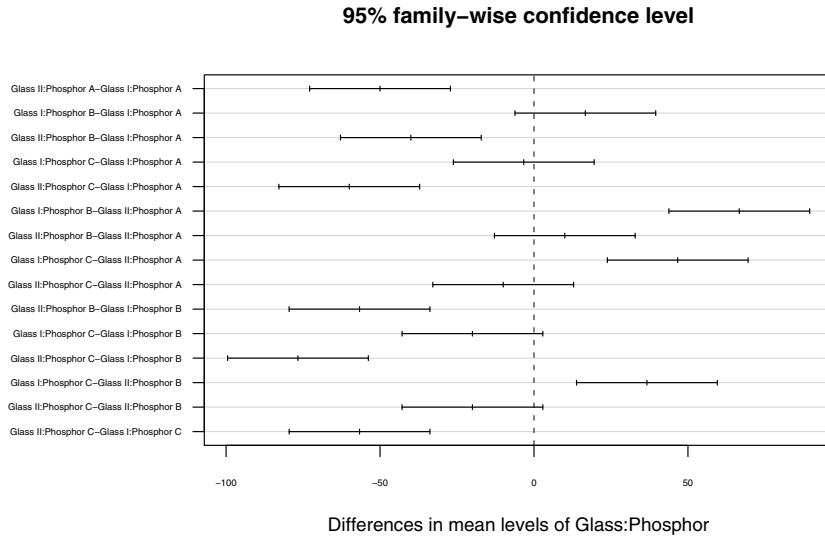


FIGURE 11.28: Tukey HSD 95% family-wise confidence intervals for the model mod.TVB

```

> library(gregmisc)
> meanM <- tapply(Microamps, list(Glass, Phosphor), mean)
> nsM <- tapply(Microamps, list(Glass, Phosphor), length)
> MSE <- anova(mod.TVB)[4, 3]
> t.c <- qt(.975, 12)
> lci <- meanM - t.c*sqrt(MSE/nsM)
> uci <- meanM + t.c*sqrt(MSE/nsM)
> barplot2(meanM, beside=TRUE, legend=TRUE, ylim=c(0, 400),
+ plot.ci=TRUE, ci.l=lci, ci.u=uci, ci.lwd=2, col=c("#A9E2FF", "#0080FF"))
> title(main="Treatment Means with Individual 95% CIs")
    
```

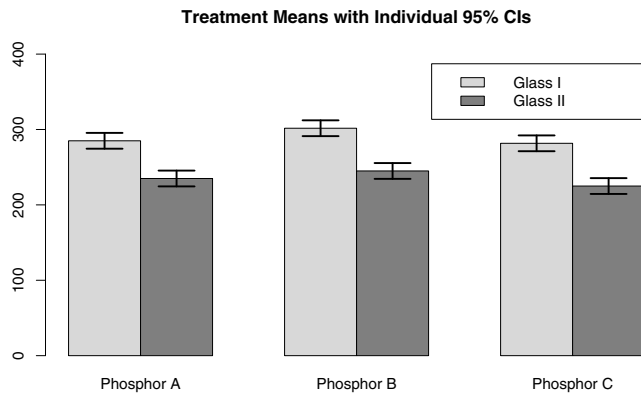


FIGURE 11.29: Barplot of the means for the six treatment combinations of factors **Glass** and **Phosphor** with individual superimposed 95% confidence intervals

11.13 Problems

1. Develop a randomization scheme to assign three treatments A, B, and C to 15 experimental units, numbered from 1 to 15. Use the command `sample` to assign them.
2. Develop a randomization scheme for a complete block design that has 4 blocks, 3 treatments, and 12 experimental units.
3. Provide a randomized assignment for a two-factor factorial design with 36 experimental units, 4 levels for the first factor, 3 levels for the second factor, and 3 experimental units for every combination of factor levels.
4. An economic study in a particular city desires to discover the monthly expenses of consumers, based on their level of education. The survey has drawn data in three different boroughs: I, II, and III. The educational levels corresponds to low, medium low, medium high, and high. The expenses have been recorded in thousands of dollars, and the analysis of variance provides the following information:

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
Ed. level	-	0.32	---	---	---
Boroughs	-	26.69	---	---	---
Residuals	-	3.64	---		

- (a) Fill in the table, and write the model corresponding to the ANOVA output.
 - (b) Calculate the percentage of the total variability explained by the educational level.
 - (c) What is the percentage of the total variability explained by the boroughs?
 - (d) What is the value for the residual variance of this model?
 - (e) Is the factor `educational level` significant?
5. Given the following partial ANOVA information from a randomized complete block design:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
block	4	0.00073	--	--	--
factor	3	--	0.35431	--	--
Residuals	--	0.01703	0.00142		

- (a) Fill in the missing values, and give the corresponding model.
 - (b) How many levels does the factor have?
 - (c) How many blocks are there in the design?
 - (d) Explain the meaning of the model's parameters.
 - (e) Describe a scenario that could be described by this particular table.
6. An agricultural engineer wants to know what type of barley produces the greatest yield: A (ASPEN), B (ERIKA), or C (SULTANE). The results obtained from 12 experimental plots, given in tons per hectare, are displayed in the following table:

A (ASPEN)	3	2	4	3
B (ERIKA)	2	3	4	4
C (SULTANE)	7	6	5	6

Assume the treatments (types of barley) were assigned at random to the 12 plots.

- (a) Write the statistical model.
 - (b) Conduct an analysis of variance and explain the results based on the model for part (a).
 - (c) Write the fitted model and provide an estimate of the error variance.
 - (d) Are the assumptions satisfied for the model in part (a)?
 - (e) Construct 95% confidence intervals for the pairwise mean differences using a technique that controls experiment-wise error.
 - (f) Use orthogonal contrasts to assess whether differences exist between SULTANE and the other two varieties ASPEN and ERIKA.
7. *Car and Driver* (July 1995) conducted tests of five cars from five different countries: Japan's Acura NSXT, Italy's Ferrari F355, Great Britain's Lotus Esprit S4S, Germany's Porche 911 Turbo, and the United States' Dodge Viper RT/10. The maximum speeds the cars obtained in miles per hour using as much distance as necessary without exceeding the engine's redline are given:

Acura	Ferrari	Lotus	Porsche	Viper
159.7	179.6	167.4	173.5	172.3
161.5	173.9	163.0	182.4	168.9
163.7	180.2	160.3	171.3	169.5
166.0	183.9	164.9	175.7	174.6
157.7	176.7	160.5	179.1	161.1
161.7	178.4	158.3	175.0	164.2

Data from Kitchens (2003).

- (a) What statistical model should be used to analyze this experiment?
 - (b) Conduct an analysis of variance to investigate if differences exist among the maximum speeds of the cars.
 - (c) Use appropriate diagnostic measures to check the adequacy of the model from part (a).
 - (d) What is the mean squared error value for the model from part (a)?
 - (e) Use Tukey's multiple comparison test to determine which of the cars are different according to speed. Plot the confidence intervals for the mean differences.
8. The data frame `barley` from the `lattice` package lists barley yield in bushels per acre for the years 1931 and 1932 for ten varieties of barley grown at six sites. Is there evidence to suggest the average barley yield in 1931 for the Waseca site is different from the average barley yield in 1931 for the Duluth site?
- (a) Use the five-step procedure to test the appropriate hypotheses using an $\alpha = 0.05$ significance level.
 - (b) Solve the same problem using a RCBD.
 - (c) Generalize your findings about the relationship between (a) and (b).
9. The following data were obtained from an experiment that investigated the effects of four bleaching chemicals (randomly selected from a large population of potential bleaching agents) on pulp brightness. The brightness of pulp is measured as the ability of a pulp sheet to reflect light directed at it. Brightness is affected by both the light

absorption and light scattering of the pulp. It is usually a measure of reflectivity, and its value is expressed as a percent of some scale (a standard measurement is the General Electric Brightness (GEB), which is a measurement of directional reflectance and is expressed as a percentage of a maximum GEB value and can be obtained by following TAPPI Standard Method T-452).

Chemical 1	77.20	74.47	72.75	76.21	72.88
Chemical 2	80.52	79.31	81.91	80.35	77.39
Chemical 3	79.42	78.02	81.60	80.80	80.63
Chemical 4	78.00	78.36	77.54	77.36	77.39

- Create side-by-side boxplots of the four chemicals. Interpret the resulting graph.
 - Specify an appropriate model to test if the chemicals have an influence on pulp brightness. Conduct an analysis on the specified model using $\alpha = 0.05$.
 - Estimate the component of variance for the chemicals.
 - Estimate the total variability in the data.
 - Construct a confidence interval for the ratio of the variability due to chemicals with respect to the total variability given that $\frac{MS_{\text{Treatment}}/(5\sigma_x^2 + \sigma^2)}{MS_{\text{Error}}/\sigma^2} \sim F_{3,21}$. Interpret your interval.
10. The household appliances section of a well-known store does research to satisfy the clients' demands for information about its products. In particular, clients are increasingly asking if the average washing times of the different brands of washing machines are the same. To discover this, the household appliances section has done the following experiment: They measured the washing time of five machines of different brands in four types of cycles (prewash, short, medium, long). The results, in minutes, are displayed in the following table:

Machines	Washing Cycle			
	Prewash	Short	Medium	Long
Machine 1	15.45	19.95	23.10	25.35
Machine 2	3.15	6.30	13.80	17.70
Machine 3	20.10	22.05	32.10	33.30
Machine 4	25.20	27.15	33.15	38.55
Machine 5	13.65	16.35	19.80	21.75

- What is the design structure used in this experiment?
- Propose a statistical model for analyzing these data.
- Use a graph to check for interaction among machines and washing cycles.
- Use diagnostic measures to check the adequacy of the model from part (b).
- If the model from part (b) is appropriate, use it to test if the average washing time is the same for the five washing machines.
- What is an estimate of the model's error variance?
- What are the estimates of the model's parameters?
- Write the model in matrix form and check that the model's constraints are satisfied.
- Use Tukey's HSD to determine which washing machines have significantly different washing times.

- (j) Is the mean washing time of machines 2, 3, and 4 significantly different from the mean washing time of machine 5?
 - (k) Use a barplot to show the mean washing times by machine. Superimpose 95% confidence intervals over the appropriate bars.
11. An insurance company wants to know how its resources are being used with respect to time spent issuing travel insurance policies. The company randomly selects three moments during a day and records the time required to issue a travel insurance policy to three randomly selected clients who take out a travel policy over the phone, over the Internet, and in person. The data obtained (in minutes) are

	1	2	3
telephone	3.49	2.38	2.09
Internet	4.38	6.68	5.37
in person	7.91	8.70	8.54

- (a) What type of design structure did the company use?
 - (b) Propose a statistical model to analyze the data.
 - (c) Comment on any assumptions that need to be made with the model selected in part (b). Check these assumptions.
 - (d) Test to see if differences exist among the methods used to issue insurance policies.
 - (e) Estimate the model's parameters.
 - (f) How is the standard deviation of the errors estimated?
 - (g) Write the estimated model in matrix form.
 - (h) Calculate the sum of residuals by rows. What can be concluded?
 - (i) Use Tukey's HSD to determine if significant differences exist among methods.
 - (j) Create a barplot of the mean times, and display the standard errors over their respective means.
12. The Environmental Protection Agency (EPA) is interested in the fuel consumption of older vehicles. An experiment is designed where the gallons of gasoline consumed by vehicles over six years old are measured when the same driver travels 162.78 miles from Boone, NC, to Durham, NC, in 35 different vehicles. Seven vehicles are randomly selected from each category to be tested. The categories are compact, station wagon, minivan, van, and full size pickup truck. The data obtained (gallons consumed) are given in the following table:

Compact	4.35	4.96	4.82	4.62	4.32	4.70	4.82
Station Wagon	5.47	6.35	5.33	6.25	5.44	5.73	5.64
Minivan	9.37	7.43	8.40	6.76	8.62	7.53	7.54
Van	8.61	8.66	10.12	8.06	9.31	6.75	8.14
Pickup Truck	20.09	14.93	13.38	16.53	13.79	12.44	14.73

- (a) Based on the described randomization, what type of design structure did the EPA use?
- (b) Propose a statistical model to analyze these data.
- (c) Are the model's assumptions specified in part (b) satisfied?

- (d) Are there significant differences between the fuel consumption for the five types of vehicles?
- (e) Estimate the model's error variance.
- (f) What conclusions can be drawn from the data?
13. A health conscious pizza parlor is attempting to specify the added calories for each ingredient of its medium size pizza. Specifically, the pizza parlor wants to know if there is more variability in an olive topping due to olive suppliers or due to the olives themselves. From numerous suppliers, four are selected randomly and the calories for a pizza topping of olives are recorded for five randomly selected pizzas. The data obtained are given in the following table:

	1	2	3	4	5
Supplier 1	133	136	142	135	134
Supplier 2	124	137	125	132	131
Supplier 3	127	126	130	120	123
Supplier 4	150	141	155	150	157

- (a) What is the design structure used in this experiment?
- (b) Specify a statistical model to analyze these data.
- (c) Conduct an ANOVA.
- (d) Estimate the variance components and the total variability of the data.
- (e) Interpret the results.
14. A turpentine manufacturer is interested in the most effective combination of acid treatment and tap hole shape for its upcoming pine resin collection. The company asks a local statistician to design an experiment to compare four tap hole shapes and to determine whether acid should be used to treat the holes. Twenty-four pine trees are selected at random from the forest where the sap will be harvested, and assigned at random to the eight combinations of acid treatment (yes or no) and hole shape (circle, diagonal slash, check, rectangle). The response is total grams of resin collected from the hole.

	Circle	Diagonal Slash	Check	Rectangle
No Acid	9	43	60	77
	13	48	65	70
	12	57	70	91
Yes Acid	15	66	75	97
	13	58	78	108
	20	73	90	99
Data in this table comes from problem 8.5, page 201 of Oehlert (2000)				

- (a) Analyze these data using a two-factor factorial design (model (11.42)).
- (b) Looking at the results of the two-way ANOVA table, is there significant interaction between acid treatment and hole shape? Use $\alpha = 0.05$.
- (c) Create a graphical display of the interactions. Does this display corroborate the numerical results?
- (d) Analyze the residuals and comment on whether model (11.42) fits the data.

- (e) If the interaction term is not significant, reanalyze a model where the interaction term is pooled with the model's error.
 - (f) Provide estimates of the parameters $\alpha_i, i = 1, 2$ and $\beta_j, j = 1, \dots, 4$. of the new model.
 - (g) Analyze the residuals.
 - (h) Use Levene's test to check if the homogeneity of variances for acid treatment levels and for hole shape levels are reasonable assumptions.
 - (i) Are the effects of acid treatment and hole shape statistically significant?
 - (j) Using an experiment-wise error rate of $\alpha_e = 0.05$, what shape has the highest quantity of resin collected?
15. The data stored in **Cows** were extracted from a Canadian record book of purebred dairy cattle. Random samples of 10 mature (five-year-old and older) and 10 two-year-old cows were taken from each of five breeds. The average butterfat percentage of these 100 cows is stored in the variable **butterfat**, with the type of cow stored in the variable **breed** and the age of the cow stored in the variable **age**.
- (a) Create a two-way ANOVA table.
 - (b) Analyze the residuals and comment on whether the two-factorial model with interaction fits the data.
 - (c) If there are problems that might be remedied with a transformation, suggest an appropriate transformation and reanalyze the new model.
 - (d) Create a graphical display of the interactions for the model selected in (b). Is there significant interaction between breed and age?
 - (e) Based on the model selected in (b), compute group means and parameter estimates to fill in a table similar to Table 11.18.
 - (f) Using $\alpha_e = 0.05$, which breed has the highest average butterfat percentage?

Case Study: Sunflower Defoliation

Ideas and data for this case study come from Muro et al. (2001).

16. Quantifying the effect of the loss of leaf area (defoliation) on sunflower (*Helianthus annuus L.*) yield caused by hail, pests, and diseases is important in the management of this crop both from a technical and economic point of view. The effect of defoliation depends, however, on the foliar surface eliminated and on the growth stage at which this takes place. The aim of this case study is to determine the response of sunflower cultivation to several levels of defoliation (**defoli**) that took place at different growth stages. An overall of 72 field trials were conducted by applying four defoliation treatments (non-defoliated control, 33%, 66% and 100%) at different growth stages (**stage**) ranging from pre-flowering (1) to physiological maturity (5) in four different locations (**location**) of Navarra, Spain: Carcastillo (1), Mélida (2), Murillo (3), and Unciti (4). There are two response variables: **yield** in kg/ha of the sunflower and **numseed**, the number of seeds per sunflower head. Data are stored in the data frame **sunflower**.
- (a) To explore the contents of the data frame **sunflower**,
 - (i) Construct a table with the total of the variable **yield** for every level of **stage** and **defoli**.

- (ii) Construct a table to display `yield` for every level of `defoli`, `location`, and `stage`. (Hint: Use the functions `xtabs()` and `ftable()`.)

Note that the function `ftable()` places 0s where there are no observations in a level combination.

- (b) How many observations are there for every combination of `stage` and `defoli`?
- (c) Is the design complete or incomplete?
- (d) Is the design balanced or unbalanced?
- (e) Use side-by-side boxplots to display the variable `yield` for every level of `stage`.
- (f) Use side-by-side boxplots to display the variable `yield` for every level of `defoli`.
- (g) Construct an interaction plot for `stage` and `defoli` on `yield`. Comment on the results.

Model (A) Conduct an analysis of variance for `yield~stage+defoli+stage:defoli`.

- (i) Is the interaction between `stage` and `defoli` statistically significant?
- (ii) Use diagnostic graphics and appropriate tests to validate this model.
- (iii) Are the assumptions for this model satisfied?

Model (B) A rogue pest infestation was found in several plots. Observations from these plots were not under experimental control. Remove any observation whose residual absolute value is greater than 2 and refit a new model.

- (i) Is the interaction term statistically significant?
- (ii) Are the assumptions of this model satisfied?

Model (C) Define a model that pools the interaction term of Model (B) with the full model's error.

- (i) Check this model's assumptions.
 - (ii) Estimate the model's effects on this model and calculate the standardized residuals. To interpret the model's effects, derive the matrix decomposition of the model as in Example 11.7.
 - (iii) Construct Tukey's HSD pairwise confidence intervals for `yield` differences by levels of `defoli` from this model. Plot the intervals and interpret the results.
 - (iv) Construct Tukey's HSD pairwise confidence intervals for `yield` differences by levels of `stage` from this model. Plot the intervals and interpret the results.
- (h) If an insurance company compensates the `yield` loss only when there is a 100% defoliation, can statistical differences between this level and the rest of the defoliation levels be found? (Hint: Use orthogonal contrasts.)
 - (i) To illustrate the final results, provide two graphs: a boxplot and a barplot of `yield` by levels of `stage`. Calculate the numerical values of the `yield` means and the standard errors.
 - (j) To illustrate the final results, redefine two levels for `defoli`: The first level combines the original levels 100 and 66 and the second new level groups the original levels 0 and 33 into a single level. Construct a boxplot and a barplot of `yield` for these two new levels. Provide a table for the corresponding means and standard errors.

Chapter 12

Regression

12.1 Introduction

The central theme of this chapter is modeling associations among variables. Understanding these associations can be important for many reasons, including:

Reason 1. Prediction of future observations

Reason 2. Variable screening

Reason 3. System explanation

Reason 4. Parameter estimation

The primary tool used to model associations among variables in this chapter is regression. Regression analysis is used for modeling the relationship between a single variable Y , called the **response** or **dependent** variable, and one or more **predictor(s)** or **independent variable(s)**, x_1, x_2, \dots, x_{p-1} . The response variable must be a continuous variable, but the predictor variables can be either continuous, discrete, or categorical. The word “regression” is due to Sir Francis Galton, who demonstrated that offspring do not tend toward the size of the parents; rather, offspring size tends toward the mean of the population. That is, there is a “*regression toward mediocrity*.” The following examples illustrate scenarios where it is important to understand the associations among response and predictor variables.

Example 12.1 ▷ *Prediction of Future Observations* ◁ A department chair is preparing a budget for the next fiscal year and must include enough money to replace personal computers in two laboratories. The chair wants to predict the price of personal computers for next year. He decides that good predictor variables for next year’s personal computer price Y are the price x_1 of a similar personal computer this year and x_2 , the rate of inflation. ■

Example 12.2 ▷ *Variable Screening* ◁ A chemist conducts a taste-testing experiment with randomly selected individuals from a particular geographical region. The dependent variable Y is the individual’s ratings of several formulations of a soft drink. The predictor variables are the various ingredients put into the soft drink. The sole purpose of the study is to decide which ingredients influence taste. ■

Example 12.3 ▷ *System Explanation* ◁ A sociologist has historical information on an isolated people group including voting records, media infiltration, numbers of roads accessing the area, and religious preferences. The sociologist is interested in understanding the rationale for why the isolated people group vote as they do. ■

Example 12.4 ▷ *Parameter Estimation* ◁ An economist has data on the GDP (gross domestic product) per capita (Y) and two independent variables: the median household income and the median household expenses for food in all European countries. The 27 points are fit to a linear model where prediction of gross domestic product is unimportant; however, the estimated signs and magnitudes of the model's parameters are important in supporting or refuting a particular economic theory. ■

Models of the form

$$Y = \beta_0 + \beta_1 x + \varepsilon \quad (12.1)$$

where β_0 and β_1 are the intercept and slope, respectively, and ε is the model error, can be used to model linear relationships between two variables. However, the population model in (12.1) is typically seen in a data setting where observations $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$ are taken on experimental units, and estimates of the parameters β_0 and β_1 are sought. In a data setting, the model is expressed as

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ for } i = 1, \dots, n \quad (12.2)$$

Model (12.2) is said to be simple, linear in the parameters (β_0 and β_1), and linear in the predictor variables (x_i). It is simple because there is only one predictor; it is linear in the parameters because no parameter appears as an exponent nor is multiplied or divided by another parameter; and it is linear in the predictor variable since the predictor variable is raised only to the first power. When the predictor variable is raised to a power, this power is called the **order** of the model.

The models

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \ln(x_i) + \varepsilon_i \\ Y_i &= \beta_0 + \beta_1 x_i x_{2i} + \varepsilon_i \end{aligned}$$

are statistical linear models; however,

$$\begin{aligned} Y_i &= \beta_0 \exp(\beta_1 x_i) + \varepsilon_i \\ Y_i &= \frac{\beta_0}{1 + e^{\beta_1 x_i}} + \varepsilon_i \end{aligned}$$

are not statistical linear models since the Y_i s are not linearly related to the parameters β_0 and β_1 . Thus, a “linear model” is characterized by a linear relationship between the dependent variable and the parameters, not necessarily by a linear relationship with the independent variables. The random error term represents the absence of an exact relationship between Y and x . When the variance for all error terms is constant, the errors are said to be homoscedastic. Typically, $\text{Var}(\varepsilon_i) = \sigma^2$. Furthermore, the random variability is independent of x . The expected value of Y given x is written

$$E[Y|x] = \beta_0 + \beta_1 x. \quad (12.3)$$

The distribution of Y given x when ε_i follows a normal distribution with a mean of zero and a standard deviation of σ is depicted in Figure 12.1 on page 566. Since the random variable Y is a linear combination of the x s, it follows that σ^2 is not truly the variance of Y but rather the variance of Y given x . As seen in Figure 12.1, $\sigma^2 = \text{Var}(\varepsilon) = \text{Var}(Y|x)$. Up to this point, normally distributed random variables have been denoted as $N(\mu, \sigma)$, where σ is the standard deviation. To simplify matrix expressions, the variance will take the place of the standard deviation in normal distributions from this point forward. For example, the distribution of the error terms in a simple linear regression model will be expressed $N(\mathbf{0}, \sigma^2 \mathbf{I})$ rather than saying each of the n errors has a $N(0, \sigma)$ distribution.

The slope, β_1 , represents the expected change in Y when a one-unit change is present in x . If $\beta_1 = 0$, Y does not depend linearly on x . When $\beta_1 < 0$, x and Y have a negative linear relationship, which means that as x increases, Y decreases. Likewise, when $\beta_1 > 0$, x and Y have a positive linear relationship, where, as x increases, so does Y .

12.2 Simple Linear Regression

The simple linear regression model when the error terms are distributed normally is

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (12.4)$$

where

Y_i is the value of the response variable for the i^{th} trial

β_0 and β_1 are parameters

x_i is a known constant for the i^{th} trial and

ε_i is a random error term that is assumed to have a $N(0, \sigma^2)$ distribution, where σ^2 (the variance) is typically unknown.

The idea that drives regression is the estimation of parameters based on n measurements. The simple linear model for n bivariate measurements is

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 x_1 + \varepsilon_1 \\ Y_2 &= \beta_0 + \beta_1 x_2 + \varepsilon_2 \\ &\vdots = \vdots + \vdots + \vdots \\ Y_n &= \beta_0 + \beta_1 x_n + \varepsilon_n \end{aligned}$$

which can also be expressed with matrix notation as

$$\mathbf{Y} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (12.5)$$

where $\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$, $\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$, $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$, and $\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$.

Note that $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$, where $\sigma^2 \mathbf{I}$ is the variance-covariance matrix of the vector of errors.

12.3 Multiple Linear Regression

Multiple linear regression is similar to simple linear regression in several ways. The dependent variable is still Y . The intercept is still β_0 . The primary change is that instead of having only β_1 as a coefficient of a single x_i variable, there now exists an entire vector

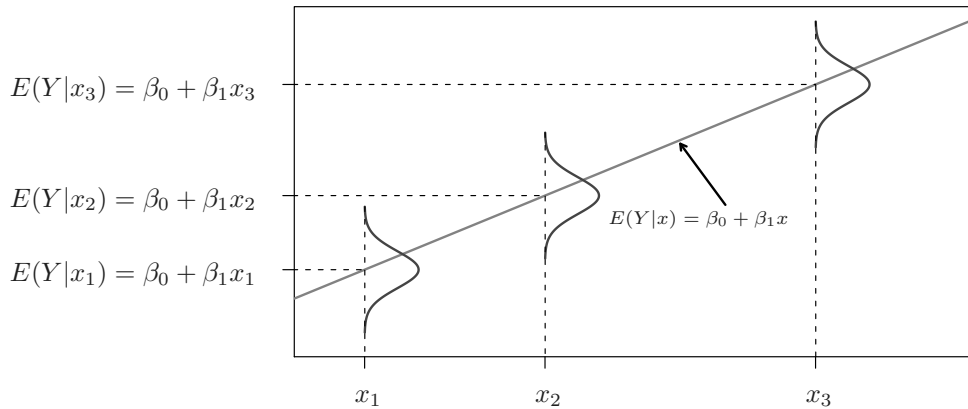


FIGURE 12.1: Graphical representation of simple linear regression model depicting the distribution of Y given x

of β_j values to multiply by a matrix of x_{ij} values. The multiple linear regression model is written

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i(p-1)} + \varepsilon_i \text{ for } i = 1, 2, \dots, n. \quad (12.6)$$

This will typically be expressed more compactly in matrix form as

$$\mathbf{Y} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (12.7)$$

where $\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$, $\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1(p-1)} \\ 1 & x_{21} & \cdots & x_{2(p-1)} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{n(p-1)} \end{bmatrix}$, $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$, and $\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$.

Each column of \mathbf{X} contains the values for a particular independent variable. The values of \mathbf{X} are assumed to be known constants. The vectors \mathbf{Y} and $\boldsymbol{\varepsilon}$ are random vectors whose elements are random variables. The vector $\boldsymbol{\beta}$ is a vector of unknown constants that are estimated from the data. Each β_j for $j = 0, 1, \dots, p-1$ indicates the change $E[Y|x_{ij}]$ for a fixed i when x_{ij} is increased by one unit and all the other predictors are held constant.

When $\boldsymbol{\varepsilon}$ is assumed $N(\mathbf{0}, \sigma^2 \mathbf{I})$, model (12.7) is referred to as the **normal error model**. In the normal error model, \mathbf{X} and $\boldsymbol{\beta}$ are assumed to be constants. Consequently, \mathbf{Y} is a random vector that is the sum of a constant vector $\mathbf{X}\boldsymbol{\beta}$ and the random vector $\boldsymbol{\varepsilon}$. Since $\boldsymbol{\varepsilon}$ is assumed $N(\mathbf{0}, \sigma^2 \mathbf{I})$, it follows that $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I})$. The tests and confidence intervals developed in later sections are based on the assumption that $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$.

Assuming that there is no error in the measurement of the x_{ij} values, one can proceed with either of the two most widely used techniques used to estimate parameters (β_j s) in a regression model: **ordinary least squares** or the **method of maximum likelihood**.

12.4 Ordinary Least Squares

The ordinary least squares method of estimating parameters minimizes the sum of the squared deviations of the Y_i s from their expected values such that

$$\varepsilon_i = Y_i - E(Y_i)$$

is the i^{th} deviation (error). For the simple linear regression model, $\varepsilon_i = Y_i - (\beta_0 + \beta_1 x_i)$. The estimates $\hat{\beta}_0$ of β_0 and $\hat{\beta}_1$ of β_1 are calculated by minimizing the quantity \mathcal{Q} (the sum of the squared residuals) found in (12.8):

$$\mathcal{Q} = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 x_i))^2, \quad (12.8)$$

which is equivalent to the matrix form

$$\mathcal{Q} = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}). \quad (12.9)$$

The values of β_0 and β_1 that minimize \mathcal{Q} are found by differentiating \mathcal{Q} with respect to β_0 and β_1 and setting the partial derivatives equal to zero. The resulting equations are known as the **normal equations**:

$$\begin{aligned} \frac{\delta \mathcal{Q}}{\delta \beta_0} &= 2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)(-1) \\ &= -2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i) \end{aligned} \quad (12.10)$$

$$\begin{aligned} \frac{\delta \mathcal{Q}}{\delta \beta_1} &= 2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)(-x_i) \\ &= -2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)(x_i) \end{aligned} \quad (12.11)$$

After setting each of these partial derivatives equal to zero, the normal equations for the simple linear regression model simplify to

$$\sum_{i=1}^n Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_i \quad (12.12)$$

$$\sum_{i=1}^n Y_i x_i = \hat{\beta}_0 \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2 \quad (12.13)$$

Note that the β_j s are replaced with $\hat{\beta}_j$ s as their values are estimates once the partial derivatives are set equal to zero. These equations are now solved for $\hat{\beta}_0$ and $\hat{\beta}_1$.

Solving for $\hat{\beta}_0$ is relatively simple using (12.12):

$$\begin{aligned}\sum_{i=1}^n Y_i &= n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_i \\ \sum_{i=1}^n Y_i - \hat{\beta}_1 \sum_{i=1}^n x_i &= n\hat{\beta}_0 \\ \frac{\sum_{i=1}^n Y_i}{n} - \frac{\hat{\beta}_1 \sum_{i=1}^n x_i}{n} &= \hat{\beta}_0 \\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{x}.\end{aligned}\tag{12.14}$$

In solving for $\hat{\beta}_1$, two quantities appear that require simplification. The first quantity is

$$\sum_{i=1}^n Y_i x_i - \frac{\sum_{i=1}^n Y_i \sum_{i=1}^n x_i}{n}.\tag{12.15}$$

$$\begin{aligned}\sum_{i=1}^n Y_i x_i - \frac{\sum_{i=1}^n Y_i \sum_{i=1}^n x_i}{n} &= \sum_{i=1}^n Y_i x_i - \bar{Y} \sum_{i=1}^n x_i \\ &= \sum_{i=1}^n Y_i x_i - \bar{Y} \sum_{i=1}^n x_i - \bar{Y} \sum_{i=1}^n x_i + \frac{n\bar{Y}}{n} \sum_{i=1}^n x_i \\ &= \sum_{i=1}^n Y_i x_i - \bar{Y} \sum_{i=1}^n x_i - \frac{\sum_{i=1}^n Y_i}{n} \sum_{i=1}^n x_i + n\bar{Y}\bar{x} \\ &= \sum_{i=1}^n Y_i x_i - \bar{Y} \sum_{i=1}^n x_i - \bar{x} \sum_{i=1}^n Y_i + n\bar{Y}\bar{x} \\ &= \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})\end{aligned}$$

The second quantity that will need to be simplified is

$$\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}.\tag{12.16}$$

$$\begin{aligned}\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} &= \sum_{i=1}^n x_i^2 - n \left(\frac{\sum_{i=1}^n x_i}{n} \right)^2 \\ &= \sum_{i=1}^n x_i^2 - n\bar{x}^2 \\ &= \sum_{i=1}^n x_i^2 - n\bar{x}^2 - n\bar{x}^2 + n\bar{x}^2 \\ &= \sum_{i=1}^n (x_i - \bar{x})^2\end{aligned}\tag{12.17}$$

Knowing these two simplifications, $\hat{\beta}_1$ can be solved using (12.13):

$$\begin{aligned}
\sum_{i=1}^n Y_i x_i &= \hat{\beta}_0 \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2 \\
\sum_{i=1}^n Y_i x_i &= (\bar{Y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2 \\
\sum_{i=1}^n Y_i x_i &= \left(\frac{\sum_{i=1}^n Y_i}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^n x_i}{n} \right) \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2 \\
\sum_{i=1}^n Y_i x_i &= \frac{\sum_{i=1}^n Y_i \sum_{i=1}^n x_i}{n} - \hat{\beta}_1 \frac{(\sum_{i=1}^n x_i)^2}{n} + \hat{\beta}_1 \sum_{i=1}^n x_i^2 \\
\sum_{i=1}^n Y_i x_i - \frac{\sum_{i=1}^n Y_i \sum_{i=1}^n x_i}{n} &= \hat{\beta}_1 \sum_{i=1}^n x_i^2 - \hat{\beta}_1 \frac{(\sum_{i=1}^n x_i)^2}{n} \\
\sum_{i=1}^n Y_i x_i - \frac{\sum_{i=1}^n Y_i \sum_{i=1}^n x_i}{n} &= \hat{\beta}_1 \left(\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} \right) \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \tag{12.18}
\end{aligned}$$

After $\hat{\beta}_0$ and $\hat{\beta}_1$ have been found, it must be shown that these values will give a minimum value for the sum of squared errors.

Proof ($\sum_{i=1}^n \hat{\varepsilon}_i^2$ is a Minimum): If the matrix of partial derivatives of \mathcal{Q} as found in (12.8) is positive definite, then our $\hat{\beta}$ values do give the minimum value for \mathcal{Q} . Recall from (12.10) that $\frac{\delta \mathcal{Q}}{\delta \beta_0} = -2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)$ and from (12.11) that $\frac{\delta \mathcal{Q}}{\delta \beta_1} = -2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)(x_i)$. This implies that the second-order partials are

$$\begin{aligned}
\frac{\delta^2 \mathcal{Q}}{\delta \beta_0^2} &= -2 \sum_{i=1}^n (-1) = 2n \\
\frac{\delta^2 \mathcal{Q}}{\delta \beta_1^2} &= -2 \sum_{i=1}^n (-x_i)(x_i) = 2 \sum_{i=1}^n x_i^2 \\
\frac{\delta^2 \mathcal{Q}}{\delta \beta_0 \delta \beta_1} &= -2 \sum_{i=1}^n (-x_i) = 2 \sum_{i=1}^n x_i
\end{aligned}$$

The matrix of partials is then

$$\frac{\delta^2 \mathcal{Q}}{\delta \beta^2} = \begin{bmatrix} 2n & 2 \sum_{i=1}^n x_i \\ 2 \sum_{i=1}^n x_i & 2 \sum_{i=1}^n x_i^2 \end{bmatrix} \tag{12.19}$$

The determinant of this matrix is $4n \sum_{i=1}^n x_i^2 - 4(\sum_{i=1}^n x_i)^2$. It must be shown that this quantity is always positive to prove that $\hat{\beta}_0$ and $\hat{\beta}_1$ as given provide a minimum value for \mathcal{Q} . Note that n is assumed to be greater than zero:

$$\begin{aligned}
4n \sum_{i=1}^n x_i^2 - 4 \left(\sum_{i=1}^n x_i \right)^2 & \stackrel{?}{>} 0 \\
\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} & \stackrel{?}{>} 0 \\
\sum_{i=1}^n (x_i - \bar{x})^2 & > 0 \quad \text{from (12.17)}
\end{aligned}$$

Therefore, the $\hat{\beta}_0$ and $\hat{\beta}_1$ calculated do give the minimum value for \mathcal{Q} . ■

Now that the β values that will minimize \mathcal{Q} are computed, the fitted regression line is written

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad (12.20)$$

where estimated (predicted) errors, also called **residuals**, are defined to be

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i. \quad (12.21)$$

12.5 Properties of the Fitted Regression Line

Several properties of the fitted regression line will be helpful in understanding the relationships between \mathbf{X} , β , ε , and \mathbf{Y} :

1. $\sum_{i=1}^n \hat{\varepsilon}_i = 0$.
2. $\sum_{i=1}^n Y_i = \sum_{i=1}^n \hat{Y}_i$.
3. $\sum_{i=1}^n x_i \hat{\varepsilon}_i = 0$.
4. $\sum_{i=1}^n \hat{Y}_i \hat{\varepsilon}_i = 0$.
5. The regression line always goes through the point (\bar{x}, \bar{Y}) .

Note that all five of these properties follow from the least squares normal (12.12) and (12.13).

Proof (Property 1):

$$\begin{aligned}
\hat{\varepsilon}_i &= Y_i - \hat{Y}_i \\
\hat{\varepsilon}_i &= Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \\
\sum_{i=1}^n \hat{\varepsilon}_i &= \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\
\sum_{i=1}^n \hat{\varepsilon}_i &= \sum_{i=1}^n Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^n x_i = 0 \quad \text{by (12.12)} \quad \blacksquare
\end{aligned}$$

Proof (Property 2):

$$\begin{aligned}\hat{\varepsilon}_i &= Y_i - \hat{Y}_i \\ \sum_{i=1}^n \hat{\varepsilon}_i &= \sum_{i=1}^n Y_i - \sum_{i=1}^n \hat{Y}_i \\ &= \sum_{i=1}^n Y_i - \sum_{i=1}^n \hat{Y}_i \\ \sum_{i=1}^n Y_i &= \sum_{i=1}^n \hat{Y}_i \quad \blacksquare\end{aligned}$$

Proof (Property 3):

$$\begin{aligned}\sum_{i=1}^n Y_i x_i &= \hat{\beta}_0 \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2 \quad \text{by (12.13)} \\ \sum_{i=1}^n Y_i x_i - \hat{\beta}_0 \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 &= 0 \\ \sum_{i=1}^n x_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) &= 0 \\ \sum_{i=1}^n x_i (Y_i - \hat{Y}_i) &= 0 \\ \sum_{i=1}^n x_i \hat{\varepsilon}_i &= 0 \quad \blacksquare\end{aligned}$$

Proof (Property 4):

$$\begin{aligned}\sum_{i=1}^n \hat{Y}_i \hat{\varepsilon}_i &= \sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 x_i) \hat{\varepsilon}_i \\ &= \hat{\beta}_0 \sum_{i=1}^n \hat{\varepsilon}_i + \hat{\beta}_1 \sum_{i=1}^n x_i \hat{\varepsilon}_i \\ &= 0 \quad \text{by Properties 1 and 3} \quad \blacksquare\end{aligned}$$

Proof (Property 5): Given the regression line $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, if $x_i = \bar{x}$, then

$$\begin{aligned}\hat{Y}_i &= \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \\ \hat{Y}_i &= \bar{Y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x} \quad \text{using (12.14)} \\ \Rightarrow \hat{Y}_i &= \bar{Y} \quad \blacksquare\end{aligned}$$

12.6 Using Matrix Notation with Ordinary Least Squares

The solutions, β , to (12.5) are generally easier to express in matrix notation than in summation notation. The normal equations are now presented in matrix form. Recall that

$$\mathcal{Q} = (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta).$$

This is simplified first and then differentiated with respect to β . Then, the result is set equal to $\mathbf{0}$ to solve for $\hat{\beta}$:

$$Q = \mathbf{Y}'\mathbf{Y} - \beta'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\beta + \beta'\mathbf{X}'\mathbf{X}\beta$$

Since $\beta'\mathbf{X}'\mathbf{Y}$ is a scalar (1×1), $(\beta'\mathbf{X}'\mathbf{Y})' = \mathbf{Y}'\mathbf{X}\beta$, so Q simplifies to

$$Q = \mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X}\beta + \beta'\mathbf{X}'\mathbf{X}\beta.$$

The expression for $\frac{\delta Q}{\delta \beta}$ can now be calculated:

$$\begin{aligned} \frac{\delta Q}{\delta \beta} &= \frac{\delta}{\delta \beta}(\mathbf{Y}'\mathbf{Y}) - \frac{\delta}{\delta \beta}(2(\mathbf{X}'\mathbf{Y})'\beta) - \frac{\delta}{\delta \beta}(\beta'\mathbf{X}'\mathbf{X}\beta) \\ &= 0 - 2\mathbf{X}'\mathbf{Y} - [\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})'\beta] \\ &\quad \text{by Rules for Differentiation 1 and 3 on page 671} \\ &= -2\mathbf{X}'\mathbf{Y} - 2\mathbf{X}'\mathbf{X}\beta \end{aligned} \tag{12.22}$$

Setting (12.22) equal to zero and solving for β yields

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, \tag{12.23}$$

the normal equations expressed in matrix notation. The worked out solutions for the matrix form of the simple linear regression model are presented next.

Recall that, for the simple linear regression model, $\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$, so

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}. \tag{12.24}$$

Also recall that the inverse of a matrix $\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is $\mathbf{A}^{-1} = \frac{1}{\det \mathbf{A}} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$, where $\det \mathbf{A} = ad - bc$. Then

$$\det(\mathbf{X}'\mathbf{X}) = n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 = n \sum_{i=1}^n (x_i - \bar{x})^2. \tag{12.25}$$

So,

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n \sum_{i=1}^n (x_i - \bar{x})^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix}. \tag{12.26}$$

Likewise,

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x_i Y_i \end{bmatrix}. \tag{12.27}$$

This means that

$$\begin{aligned}
\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\
&= \frac{1}{n \sum_{i=1}^n (x_i - \bar{x})^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x_i Y_i \end{bmatrix} \\
&= \begin{bmatrix} \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n Y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i Y_i}{n \sum_{i=1}^n (x_i - \bar{x})^2} \\ \frac{-\sum_{i=1}^n x_i \sum_{i=1}^n Y_i + n \sum_{i=1}^n x_i Y_i}{n \sum_{i=1}^n (x_i - \bar{x})^2} \end{bmatrix} \tag{12.28} \\
&= \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}
\end{aligned}$$

Next, it should be shown that the matrix solutions for $\hat{\beta}_0$ and $\hat{\beta}_1$ are identical to the summation solutions shown in (12.14) and (12.18). Converting the second entry in $\hat{\beta}$ from (12.28) to (12.18) is more obvious, so it will be done first:

$$\begin{aligned}
&\frac{-\sum_{i=1}^n x_i \sum_{i=1}^n Y_i + n \sum_{i=1}^n x_i Y_i}{n \sum_{i=1}^n (x_i - \bar{x})^2} \stackrel{?}{=} \hat{\beta}_1 \\
&\frac{\frac{1}{n} \cdot n \sum_{i=1}^n x_i Y_i - \sum_{i=1}^n x_i \sum_{i=1}^n Y_i}{n \sum_{i=1}^n (x_i - \bar{x})^2} \stackrel{?}{=} \hat{\beta}_1 \\
&\frac{\sum_{i=1}^n x_i Y_i - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n Y_i}{n}}{\sum_{i=1}^n (x_i - \bar{x})^2} \stackrel{?}{=} \hat{\beta}_1 \\
&\frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \hat{\beta}_1 \text{ by simplification (12.15)}
\end{aligned}$$

Next, show $\hat{\beta}_0$ from (12.14) is equal to the first entry of $\hat{\beta}$:

$$\begin{aligned}
&\frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n Y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i Y_i}{n \sum_{i=1}^n (x_i - \bar{x})^2} \stackrel{?}{=} \hat{\beta}_0 \\
&\frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n Y_i - \frac{\left(\sum_{i=1}^n x_i\right)^2 \sum_{i=1}^n Y_i}{n} + \frac{\left(\sum_{i=1}^n x_i\right)^2 \sum_{i=1}^n Y_i}{n} - \sum_{i=1}^n x_i \sum_{i=1}^n x_i Y_i}{n \sum_{i=1}^n (x_i - \bar{x})^2} \stackrel{?}{=} \hat{\beta}_0 \\
&\frac{\sum_{i=1}^n Y_i \left[\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} \right] - \left[\left(\sum_{i=1}^n x_i Y_i - \bar{Y} \sum_{i=1}^n x_i \right) \sum_{i=1}^n x_i \right]}{n \sum_{i=1}^n (x_i - \bar{x})^2} \stackrel{?}{=} \hat{\beta}_0 \\
&\frac{\sum_{i=1}^n Y_i [\sum_{i=1}^n (x_i - \bar{x})^2] - [\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}) \sum_{i=1}^n x_i]}{n \sum_{i=1}^n (x_i - \bar{x})^2} \stackrel{?}{=} \hat{\beta}_0 \\
&\text{by Simplifications (12.16) and (12.15)} \\
&\frac{\sum_{i=1}^n Y_i [\sum_{i=1}^n (x_i - \bar{x})^2]}{n \sum_{i=1}^n (x_i - \bar{x})^2} - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{n \sum_{i=1}^n (x_i - \bar{x})^2} \stackrel{?}{=} \hat{\beta}_0 \\
&\bar{Y} - \hat{\beta}_1 \bar{x} = \hat{\beta}_0
\end{aligned}$$

Therefore, the matrix solution is identical to the summation solution, so $\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$.

Example 12.5 Find the variance-covariance matrix of $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ when \mathbf{X} is an $n \times p$ matrix.

Solution: Let $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{A}$. Then $\hat{\beta} = \mathbf{A}\mathbf{Y}$ and $\sigma_{\hat{\beta}}^2 = \mathbf{A}\sigma_{\mathbf{Y}}^2\mathbf{A}'$ by property 3 on page 673. Note that $\sigma_{\mathbf{Y}}^2 = \sigma^2\mathbf{I}_{n \times n}$. So,

$$\begin{aligned}\sigma_{\hat{\beta}}^2 &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}_{n \times n}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

■

Example 12.6 ▷ *Linear Relationship between GPA and SAT Scores* ◁ The admissions committee of a comprehensive state university selected at random the records of 200 second-semester freshmen. The results, first-semester college GPA and SAT scores, are stored in the data frame **Grades**. The admissions committee wants to study the linear relationship between first-semester college grade point average (**gpa**) and scholastic aptitude test (**sat**) scores. Assume that the requirements for model (12.4) are satisfied.

- Create a scatterplot of the data to investigate the relationship between **gpa** and **sat** scores.
- Obtain the least squares estimates for β_0 and β_1 , and state the estimated regression function using
 - Summation notation with (12.14) and (12.18).
 - Matrix notation with (12.23).
 - Use the S function `lm()` to verify the answers in (i) and (ii).
- What is the point estimate of the change in the mean **gpa** when the **sat** score increases by 50 points?

Solution: The data frame **Grades** is in the PASWR package.

- The scatterplot in Figure 12.2 on the next page suggests a linear relationship exists between **gpa** and **sat**.

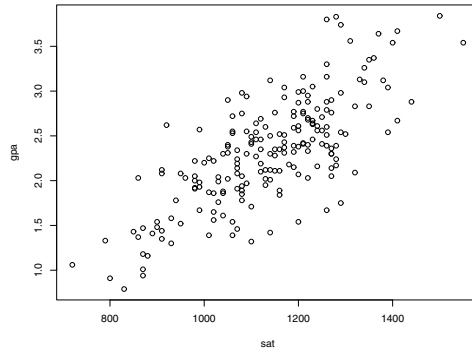
```
> attach(Grades)
> plot(sat, gpa)
```

(b)

- Assign **gpa** to Y and **sat** to x :

```
> Y <- gpa
> x <- sat
```

Solving using summation notation as in (12.18) (b1) and (12.14) (b0):

FIGURE 12.2: Scatterplot of `gpa` versus `sat` using `Grades`

```
> b1 <- sum((x-mean(x))*(Y-mean(Y))) / sum((x-mean(x))^2)
> b0 <- mean(Y)-b1*mean(x)
> c(b0, b1)
[1] -1.192063812  0.003094270
```

The estimated regression function is $\hat{Y}_i = -1.192063812 + 0.003094270x_i$.

(ii) Solving using matrix notation as in (12.23):

```
> X <- cbind(rep(1,200), x)
> Y <- matrix(Y, nrow=200)
> betahat <- solve(t(X)%*%X)%*%t(X)%*%Y
> beta0hat <- betahat[1,1]
> beta1hat <- betahat[2,1]
> c(beta0hat, beta1hat)
[1] -1.192063812  0.003094270
```

The estimated regression function is $\hat{Y}_i = -1.192063812 + 0.003094270x_i$.

(iii) Solving use `lm()`:

```
> model <- lm(Y~x)
> model$coefficients
(Intercept)          x
-1.192063812  0.003094270
```

The estimated regression function is $\hat{Y}_i = -1.192063812 + 0.003094270x_i$.

(c) The point estimate of the change in the mean `gpa` when the SAT score increases by 50 points is $\hat{\beta}_1 \cdot 50 = 0.1547135$:

```
> b1*50
[1] 0.1547135
> detach(Grades)
```



12.7 The Method of Maximum Likelihood

The method of least squares is not the only one that can be used to construct an estimate of β . Another common method for constructing estimators is that of maximum likelihood. To construct the maximum likelihood estimator (MLE) of β when $\varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$, first construct the likelihood function. Next, take the natural log. Finally, take appropriate partial derivatives and set them equal to zero to solve for the MLE of β , $\tilde{\beta}$.

The likelihood function for β and σ^2 when \mathbf{X} is given is

$$\mathcal{L}(\beta, \sigma^2 | \mathbf{X}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[\frac{-(Y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1}))^2}{2\sigma^2} \right] \quad (12.29)$$

In matrix form, this is

$$\mathcal{L}(\beta, \sigma^2 | \mathbf{X}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[\frac{-(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)}{2\sigma^2} \right] \quad (12.30)$$

The natural log of the matrix form of the likelihood function (log-likelihood function) is

$$\ln \mathcal{L}(\beta, \sigma^2 | \mathbf{X}) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)}{2\sigma^2}$$

Simplifying the partial derivative of the log-likelihood function with respect to β gives

$$\begin{aligned} \frac{\delta \ln \mathcal{L}(\beta, \sigma^2 | \mathbf{X})}{\delta \beta} &= \frac{\delta}{\delta \beta} \left[-\frac{(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)}{2\sigma^2} \right] \\ &= \frac{\delta}{\delta \beta} \left[\frac{-\mathbf{Y}'\mathbf{Y} + \beta'\mathbf{X}'\mathbf{Y} + \mathbf{Y}'\mathbf{X}\beta - \beta'\mathbf{X}'\mathbf{X}\beta}{2\sigma^2} \right] \end{aligned}$$

Recall that $\beta'\mathbf{X}'\mathbf{Y}$ is 1×1

$$= \frac{\delta}{\delta \beta} \left[\frac{-\mathbf{Y}'\mathbf{Y} + 2(\mathbf{X}'\mathbf{Y})'\beta - \beta'\mathbf{X}'\mathbf{X}\beta}{2\sigma^2} \right]$$

By Rules of Differentiation 1 and 3 on page 671

$$= \frac{2\mathbf{X}'\mathbf{Y}}{2\sigma^2} - \left[\frac{\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})'\beta}{2\sigma^2} \right]$$

$$= \frac{\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\beta}{\sigma^2}$$

Setting this equal to zero and solving for β yields

$$\tilde{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Note that the MLE is equivalent to the ordinary least squares estimator for β given in (12.23). It is also of interest to find the MLE for σ^2 . Taking the partial derivative of the log-likelihood function in terms of σ^2 gives

$$\frac{\delta \ln \mathcal{L}(\beta, \sigma^2 | \mathbf{X})}{\delta \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \cdot (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)$$

When this quantity is set equal to zero and solved for σ^2 , the MLE is

$$\tilde{\sigma}^2 = \frac{(\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta})}{n} = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n}$$

Unfortunately, $\tilde{\sigma}^2$ is a biased estimator of σ^2 . The bias is easily fixed and the unbiased estimator $\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-p}$ is typically used to estimate σ^2 .

12.8 The Sampling Distribution of $\hat{\beta}$

The matrix form of $\hat{\beta}$ was described in (12.23) to be $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. If model (12.7) assumes that $\varepsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, then $\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$ because \mathbf{X} and β are assumed to be constants. Since $\hat{\beta}$ can be expressed as constants multiplied by \mathbf{Y} , it follows that $\hat{\beta}$ also has a normal distribution. It will be shown that

$$\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}).$$

In Example 12.5 on page 574, the variance of $\hat{\beta}$ was shown to equal $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Next, $\hat{\beta}$ is shown to be an unbiased estimator of β . Specifically,

$$\begin{aligned} \text{If } \hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ \text{Then } E[\hat{\beta}] &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \varepsilon)] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon] \\ &= E[\mathbf{I}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon] \\ &= \beta \text{ since } \mathbf{I} \text{ and } (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \text{ are constants and } E(\varepsilon) = \mathbf{0}. \end{aligned}$$

under the normal error regression model. However, unbiasedness does not guarantee uniqueness. Fortunately, the **Gauss-Markov** theorem guarantees that among the class of linear unbiased estimators for β , $\hat{\beta}$ is the best in the sense that the variances of $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ are minimized. Consequently, $\hat{\beta}$ is called a best linear unbiased estimator, or a **BLUE**. Note that the error variance σ^2 is unknown, but its unbiased estimate is given by

$$\hat{\sigma}^2 = s^2 = MSE = \frac{SSE}{n-p} = \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{n-p} \quad (12.31)$$

If the matrix \mathbf{V} is defined to be $(\mathbf{X}'\mathbf{X})^{-1}$, then $\sigma_{\hat{\beta}_k}^2 = \sigma^2 \cdot v_{k+1, k+1}$, where $v_{k+1, k+1}$ is the $(k+1)^{\text{st}}$ diagonal entry ($k = 0, 1, \dots, p-1$) of \mathbf{V} . It is preferable to calculate \mathbf{V} with the command `summary(lm.object)$cov.unscaled`, where `lm.object` is a linear model object, rather than with the matrix computations `t(X)%*%X`, where `X` is the design matrix. Since $\hat{\beta} \sim N(\beta, \sigma_{\hat{\beta}}^2)$, where $\sigma_{\hat{\beta}}^2 = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, an estimate of $\sigma_{\hat{\beta}}^2$ is

$$\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1} = MSE(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} s_{\hat{\beta}_0}^2 & s_{\hat{\beta}_0, \hat{\beta}_1} & \cdots & s_{\hat{\beta}_0, \hat{\beta}_{p-1}} \\ s_{\hat{\beta}_1, \hat{\beta}_0} & s_{\hat{\beta}_1}^2 & \cdots & s_{\hat{\beta}_1, \hat{\beta}_{p-1}} \\ \vdots & \vdots & \ddots & \vdots \\ s_{\hat{\beta}_{p-1}, \hat{\beta}_0} & s_{\hat{\beta}_{p-1}, \hat{\beta}_1} & \cdots & s_{\hat{\beta}_{p-1}}^2 \end{bmatrix} = \mathbf{s}_{\hat{\beta}}^2. \quad (12.32)$$

The R function `vcov()` will compute $\mathbf{s}_{\hat{\beta}}^2$ when applied to a linear model object. A test statistic for testing $H_0 : \beta_k = \beta_{k_0}$ versus $H_0 : \beta_k \neq \beta_{k_0}$ can be justified using the standard form of a t -statistic:

$$\frac{\text{unbiased estimator} - \text{hypothesized value}}{\text{standard error of estimator}}$$

Specifically, the test statistic is

$$\frac{\hat{\beta}_k - \beta_{k_0}}{s_{\hat{\beta}_k}} \sim t_{n-p} \text{ for } k = 0, 1, \dots, p-1 \quad (12.33)$$

In the event that the hypothesis of interest is $H_0 : \beta_k = 0$ versus $H_1 : \beta_k \neq 0$, the function `summary()` applied to a linear model object will provide the $t_{\text{obs}} = \hat{\beta}_k / s_{\hat{\beta}_k}$ value and the corresponding φ -value. That is, the φ -value $= 2 \times \mathbb{P}(t_{n-p} \geq |t_{\text{obs}}|)$.

Using (12.33) as a pivotal quantity, in a similar fashion to the derivation of a confidence interval for μ in Section 8.2.2, a $100 \cdot (1 - \alpha)\%$ confidence interval for β_k , where $k = 0, 1, \dots, p - 1$, is

$$CI_{1-\alpha}(\beta_k) = \left[\hat{\beta}_k - t_{1-\alpha/2; n-p} \cdot s_{\hat{\beta}_k}, \hat{\beta}_k + t_{1-\alpha/2; n-p} \cdot s_{\hat{\beta}_k} \right]. \quad (12.34)$$

Note that the degrees of freedom for the t -distribution are $n - p$ because σ^2 is estimated with $MSE = \frac{SSE}{n-p}$.

Example 12.7 Consider Example 12.6 on page 574, where the admissions committee of a comprehensive state university wants to study the linear relationship between first-semester college grade point averages (`gpa`) and scholastic aptitude test (`sat`) scores. These are stored in the data frame `Grades`. Assume that the requirements for model (12.4) are satisfied.

- Find the variance-covariance matrix for $\hat{\beta}$ using (12.32).
- Test whether there is a linear relationship at the $\alpha = 0.10$ significance level.
- Construct 90% confidence intervals for β_0 and β_1 .

Solution: (a) Recall that $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$, $\hat{\sigma}^2 = \sum_{i=1}^n \frac{\hat{\varepsilon}_i^2}{n-p}$, and the variance-covariance matrix is $s_{\hat{\beta}}^2 = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$:

```
> attach(Grades)
> Y <- gpa
> x <- sat
> simple.model <- lm(Y~x)
> X <- cbind(rep(1,200), x)
> XTX <- t(X)%*%X
> solve(XTX)
              x
0.3101379642 -2.689270e-04
x -0.0002689270  2.370131e-07
```

The preferred way to compute $(\mathbf{X}'\mathbf{X})^{-1}$ with S is

```
> XTXI <- summary(simple.model)$cov.unscaled
> XTXI
              (Intercept)              x
(Intercept) 0.3101379642 -2.689270e-04
x           -0.0002689270  2.370131e-07
> MSE <- sum(resid(simple.model)^2)/(200-2)
> MSE
[1] 0.1595551
```

A more direct method of obtaining the MSE is `summary(simple.model)$sigma^2`:

```
> var.cov.b <- MSE*XTXI
> var.cov.b
      (Intercept)          x
(Intercept) 4.948408e-02 -4.290866e-05
x           -4.290866e-05  3.781665e-08
```

$$s_{\hat{\beta}}^2 = \begin{bmatrix} 4.948408 \times 10^{-2} & -4.290866 \times 10^{-5} \\ -4.290866 \times 10^{-5} & 3.781665 \times 10^{-8} \end{bmatrix}$$

To compute $s_{\hat{\beta}}^2$ with R directly, type

```
> vcov(simple.model)
      (Intercept)          x
(Intercept) 4.948408e-02 -4.290866e-05
x           -4.290866e-05  3.781665e-08
```

(b)

Step 1: **Hypotheses** — $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$.

Step 2: **Test Statistic** — $\hat{\beta}_1 = 0.0030943$ is the test statistic. Assuming the assumptions of Model (12.4) are satisfied,

$$\hat{\beta}_1 \sim N(\beta_1, \sigma_{\hat{\beta}_1}^2)$$

The standardized test statistic under the assumption that H_0 is true and its distribution are

$$\frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} \sim t_{200-2}$$

Step 3: **Rejection Region Calculations** — Because the standardized test statistic is distributed t_{198} and H_1 is a two-sided hypothesis, the rejection region is $|t_{\text{obs}}| > t_{0.95;198} = 1.6526$. The value of the standardized test statistic is $t_{\text{obs}} = \frac{0.0031-0}{.00019} = 15.912$.

Step 4: **Statistical Conclusion** — The φ -value is $2 \times \mathbb{P}(t_{198} \geq 15.912) = 2 \times 0 = 0$.

I. From the rejection region, reject H_0 because $|15.912|$ is greater than 1.6526.

II. From the φ -value, reject H_0 because the φ -value = 0 is less than 0.10.

Step 5: **English Conclusion** — There is evidence to suggest a linear relationship between sat and gpa.

To see the test statistics and their φ -values for the `simple.model`, enter

```
> summary(simple.model)$coef
      Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.192063812 0.2224501802 -5.35879 2.316666e-07
x           0.003094270 0.0001944650 15.91171 2.922995e-37
```

(c) 90% Confidence intervals for β_0 and β_1 are

$$CI_{0.90}(\beta_0) = \left[\hat{\beta}_0 - t_{.95; n-p} \cdot s_{\hat{\beta}_0}, \hat{\beta}_0 + t_{.95; n-p} \cdot s_{\hat{\beta}_0} \right]$$

$$CI_{0.90}(\beta_0) = [-1.19 - 1.65(0.22), -1.19 + 1.65(0.22)]$$

$$CI_{0.90}(\beta_0) = [-1.56, -0.82]$$

and

$$CI_{0.90}(\beta_1) = \left[\hat{\beta}_1 - t_{.95; n-p} \cdot s_{\hat{\beta}_1}, \hat{\beta}_1 + t_{.95; n-p} \cdot s_{\hat{\beta}_1} \right]$$

$$CI_{0.90}(\beta_1) = [0.003 - 1.65(0.00019), 0.003 + 1.65(0.00019)]$$

$$CI_{0.90}(\beta_1) = [0.0028, 0.0034].$$

```
> b0 <- summary(simple.model)$coef[1,1]
> b1 <- summary(simple.model)$coef[2,1]
> s.b0 <- summary(simple.model)$coef[1,2]
> s.b1 <- summary(simple.model)$coef[2,2]
> ct <- qt(1-.10/2,198) # alpha = 0.10
> ct
[1] 1.652586
> CI.B0 <- c(b0 - ct*s.b0, b0 + ct*s.b0)
> CI.B0
[1] -1.5596818 -0.8244458
> CI.B1 <- c(b1 - ct*s.b1, b1 + ct*s.b1)
> CI.B1
[1] 0.002772900 0.003415640
```

Or, if working in R only, a method requiring less typing is

```
> confint(lm(Y~x), level=.90)
          5 %          95 %
(Intercept) -1.5596818 -0.824445807
x             0.0027729  0.003415640
> detach(Grades)
```

■

12.9 ANOVA Approach to Regression

The basic normal error term regression model (12.4) has now been developed extensively. This same model can also be considered in an analysis of variance framework. This new paradigm will prove useful when working with multiple regression models. The analysis of variance approach is based on partitioning the sums of squares and the degrees of freedom associated with the response variable Y . The total deviation, $Y_i - \bar{Y}$, can be decomposed into two components:

1. The deviation of the fitted value \hat{Y}_i around the mean \bar{Y} and

2. The deviation of the observation Y_i around the regression line.

$$\underbrace{Y_i - \bar{Y}}_{\text{Total Deviation}} = \underbrace{\hat{Y}_i - \bar{Y}}_{\substack{\text{Deviation of Fitted} \\ \text{Regression Value} \\ \text{around the Mean}}} + \underbrace{Y_i - \hat{Y}_i}_{\text{Deviation around the Fitted Regression Line}} \quad (12.35)$$

Note that the total deviation is used to measure the variation of the Y_i s without taking the predictor variable(s) into account. Recall that since $\hat{\epsilon}_i = Y_i - \hat{Y}_i$,

$$\begin{aligned} 2 \sum_{i=1}^n (\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) &= 2 \sum_{i=1}^n (\hat{Y}_i - \bar{Y}) \hat{\epsilon}_i \\ &= 2 \sum_{i=1}^n \hat{Y}_i \hat{\epsilon}_i - 2\bar{Y} \sum_{i=1}^n \hat{\epsilon}_i \\ &= \underbrace{2 \times 0}_{\text{by Property (4)}} - \underbrace{2 \times 0}_{\text{by Property (1)}} = 0, \end{aligned} \quad (12.36)$$

so that squaring both sides and summing from $i = 1$ to n of (12.35) yields

$$\begin{aligned} \sum_{i=1}^n (Y_i - \bar{Y})^2 &= \sum_{i=1}^n \left[(\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i) \right]^2 \\ &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \underbrace{2 \sum_{i=1}^n (\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i)}_{=0 \text{ by (12.36)}} \\ \sum_{i=1}^n (Y_i - \bar{Y})^2 &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \end{aligned} \quad (12.37)$$

The expression in (12.37) is commonly expressed as $SST = SSR + SSE$, where SST denotes total sum of squares, SSR stands for regression sum of squares, and SSE represents error (residual) sum of squares.

12.9.1 ANOVA with Simple Linear Regression

The degrees of freedom for SST are partitioned into degrees of freedom for SSR and degrees of freedom for SSE , just as the total sum of squares (SST) itself was partitioned into SSR and SSE . There are $n - 1$ degrees of freedom associated with SST . One degree of freedom is lost since the deviations $Y_i - \bar{Y}$ are subject to one constraint, specifically, $\sum_{i=1}^n (Y_i - \bar{Y})$ must equal zero, as it always does. Another explanation is that one degree of freedom is lost since \bar{Y} is used to estimate the population mean, μ .

SSE has $n - 2$ degrees of freedom. Two degrees of freedom are lost since two parameters, β_0 and β_1 , are estimated while obtaining the fitted values of \hat{Y}_i . There are two degrees of freedom associated with the estimated regression line, that is, one for the slope and one for the intercept. However, one of the degrees of freedom is lost since $\sum_{i=1}^n (\hat{Y}_i - \bar{Y})$ must equal zero by property 2. Consequently, SSR in a simple linear regression model has one degree of freedom.

When a sum of squares is divided by its associated degrees of freedom, the result is called a **mean square** and is denoted with MS . Specifically,

$$\frac{SSR}{1} = MSR \text{ and } \frac{SSE}{n - 2} = MSE$$

Mean squares, unlike sums of squares, are not additive. That is,

$$MST \neq MSR + MSE.$$

For the normal error regression model in (12.4), $\frac{SSE}{\sigma^2} \sim \chi_{n-2}^2$. Consequently,

$$E\left[\frac{SSE}{\sigma^2}\right] = n - 2 \implies E\left[\frac{SSE}{n - 2}\right] = \sigma^2 \implies E[MSE] = \sigma^2$$

In other words, the MSE is an unbiased estimator of σ^2 .

To find the expected value of MSR , recall from property 5 that $\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$ and that the SSR for the simple linear model has one degree of freedom. This implies that $SSR = SSR/1 = MSR$. Also, note that the definition of the variance of $\hat{\beta}_1$ is $\sigma_{\hat{\beta}_1}^2 = E[\hat{\beta}_1^2] - (E[\hat{\beta}_1])^2 \Rightarrow E[\hat{\beta}_1^2] = \sigma_{\hat{\beta}_1}^2 + (E[\hat{\beta}_1])^2$.

$$\begin{aligned} SSR &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \\ SSR &= \sum_{i=1}^n ((\hat{\beta}_0 + \hat{\beta}_1 x_i) - (\hat{\beta}_0 + \hat{\beta}_1 \bar{x}))^2 \\ SSR &= \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 \end{aligned}$$

Then, $E[SSR] = E[\hat{\beta}_1^2] \sum_{i=1}^n (x_i - \bar{x})^2$, since the x values are not random:

$$\begin{aligned} E[SSR] &= \left\{ \sigma_{\hat{\beta}_1}^2 + (E[\hat{\beta}_1])^2 \right\} \cdot \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \left\{ \frac{\sigma^2}{\underbrace{\sum_{i=1}^n (x_i - \bar{x})^2}_{\text{by Example 12.5 and (12.26)}}} + \beta_1^2 \right\} \cdot \sum_{i=1}^n (x_i - \bar{x})^2 \end{aligned}$$

$$E[SSR] = \sigma^2 + \beta_1^2 \cdot \sum_{i=1}^n (x_i - \bar{x})^2 = E[SSR/1] =$$

$$E[MSR] = \sigma^2 + \beta_1^2 \cdot \sum_{i=1}^n (x_i - \bar{x})^2$$

Note that the mean of the sampling distribution of MSE is σ^2 whether a linear relationship exists between Y and x or not. The mean of the sampling distribution of MSR is also σ^2 when $\beta_1 = 0$. Consequently, MSR and MSE will be similar in magnitude when $\beta_1 = 0$. Likewise, when $\beta_1 \neq 0$, the center of the sampling distribution of MSR will be larger than the center of the sampling distribution of MSE by approximately $\beta_1^2 \cdot \sum_{i=1}^n (x_i - \bar{x})^2$.

In particular, the test statistic for testing $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ for model (12.4) is

$$F_{\text{obs}} = \frac{MSR}{MSE}. \quad (12.38)$$

When the null hypothesis is true, $H_0 : \beta_1 = 0$, then

$$\frac{MSR}{MSE} \sim F_{1, n-2}.$$

Although it is beyond the scope of this text, it is noted that the quantities $\frac{SSR}{\sigma^2}$ and $\frac{SSE}{\sigma^2}$ are independent χ^2 random variables with 1 and $n - 2$ degrees of freedom, respectively. It then follows, using Definition 6.2, that

$$\frac{MSR}{MSE} = \frac{\frac{SSR/\sigma^2}{1}}{\frac{SSE/\sigma^2}{n-2}} = \frac{\chi_1^2/1}{\chi_{n-2}^2/(n-2)} \sim F_{1,n-2}.$$

Finally, values of F_{obs} close to 1 tend to support the null hypothesis, while large values of F_{obs} tend to support the alternative hypothesis. Specifically, the null hypothesis is rejected if $F_{\text{obs}} > f_{1-\alpha;1,n-2}$. `S` generates an ANOVA table on linear model objects with the function `anova(lm.object)`.

Table 12.1: ANOVA table for simple linear regression

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	F_{obs}
Regression	1	$SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	$MSR = \frac{SSR}{1}$	$\frac{MSR}{MSE}$
Error	$n - 2$	$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	$MSE = \frac{SSE}{n-2}$	
Total	$n - 1$	$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$		

Example 12.8 Construct an ANOVA table using the data in `Grades`. Then, test if a linear relationship exists between first-semester college grade point average (`gpa`) and scholastic aptitude score (`sat`) using the information in the ANOVA table at the $\alpha = 0.05$ level.

Solution: The following code is used to create Table 12.2:

```
> attach(Grades)
> model.lm <- lm(gpa~sat)
> anova(model.lm)
> detach(Grades)
```

Table 12.2: ANOVA table for `model.lm <- lm(gpa~sat)`

anova	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sat	1	40.397	40.397	253.18	$< 2.2e - 16$
Residuals	198	31.592	0.160		

Step 1: **Hypotheses** — $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$.

Step 2: **Test Statistic** — F_{obs} since MSR/MSE under the assumption that $\beta_1 = 0$ has an $F_{1,198}$ distribution.

Step 3: **Rejection Region Calculations** — Because $F_{\text{obs}} \sim F_{1,198}$ and this is a one-tailed test, the rejection region is $F_{\text{obs}} > f_{0.95;1,198} = 3.888$. The value of the standardized test statistic is $F_{\text{obs}} = \frac{40.40}{0.16} = 253.18$.

Step 4: **Statistical Conclusion** — The ϕ -value is $\mathbb{P}(F_{1,198} \geq 253.18) = 0$.

- I. From the rejection region, reject H_0 because 253.18 is greater than 3.888.
- II. From the ϕ -value, reject H_0 because the ϕ -value = 0 is less than 0.05.

Step 5: **English Conclusion** — There is strong evidence to suggest a linear relationship exists between first-semester `gpa` and `sat` scores. ■

12.9.2 ANOVA with Multiple Linear Regression

The ANOVA approach to linear regression analysis can be generalized to test hypotheses of the form

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0 \text{ versus} \\ H_1 : \text{at least one } \beta_i \neq 0 \text{ for } i = 1, 2, \dots, p-1$$

The ANOVA table for a multiple linear regression model with p parameters expressed in matrix notation is given in Table 12.3. Note that \mathbf{J} is an $n \times n$ matrix of 1s.

Table 12.3: ANOVA table for multiple linear regression

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	F_{obs}
Regression	$p - 1$	$SSR = \hat{\beta} \mathbf{X}' \mathbf{Y} - \frac{1}{n} \mathbf{Y}' \mathbf{J} \mathbf{Y}$	$MSR = \frac{SSR}{p-1}$	$\frac{MSR}{MSE}$
Error	$n - p$	$SSE = \mathbf{Y}' \mathbf{Y} - \hat{\beta}' \mathbf{X}' \mathbf{Y}$	$MSE = \frac{SSE}{n-p}$	
Total	$n - 1$	$SST = \mathbf{Y}' \mathbf{Y} - \frac{1}{n} \mathbf{Y}' \mathbf{J} \mathbf{Y}$		

An important matrix in the theory of linear models is the \mathbf{H} or “hat” matrix, defined as

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \quad (12.39)$$

The \mathbf{H} matrix is a symmetric, idempotent ($\mathbf{H}^2 = \mathbf{H}$), $n \times n$ matrix that transforms the Y_i s into \hat{Y}_i s. Specifically,

$$\begin{aligned} \hat{\mathbf{Y}} &= \mathbf{X}\hat{\beta} \\ \hat{\mathbf{Y}} &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ \hat{\mathbf{Y}} &= \mathbf{H}\mathbf{Y} \end{aligned}$$

The values for the sums of squares found in Table 12.3 can also be expressed in terms of the hat matrix as well as identity and \mathbf{J} matrices. Recall that $\hat{\beta}'\mathbf{X}'\mathbf{Y}$ is a 1×1 vector and is thus equal to its transpose:

$$\begin{aligned}
SSE &= \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} \\
&= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\beta}} \\
&= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\
&= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{H}\mathbf{Y} \\
&= \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y} \quad (12.40)
\end{aligned}$$

$$\begin{aligned}
SSR &= \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} - \frac{1}{n}\mathbf{Y}'\mathbf{J}\mathbf{Y} \\
&= \mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\beta}} - \frac{1}{n}\mathbf{Y}'\mathbf{J}\mathbf{Y} \\
&= \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} - \frac{1}{n}\mathbf{Y}'\mathbf{J}\mathbf{Y} \\
&= \mathbf{Y}'\mathbf{H}\mathbf{Y} - \frac{1}{n}\mathbf{Y}'\mathbf{J}\mathbf{Y} \\
&= \mathbf{Y}'(\mathbf{H} - \frac{1}{n}\mathbf{J})\mathbf{Y} \quad (12.42)
\end{aligned}$$

$$\begin{aligned}
SST &= \mathbf{Y}'\mathbf{Y} - \frac{1}{n}\mathbf{Y}'\mathbf{J}\mathbf{Y} \\
&= \mathbf{Y}'(\mathbf{I} - \frac{1}{n}\mathbf{J})\mathbf{Y} \quad (12.41)
\end{aligned}$$

Thus it can be seen that each of the three sums of squares can be expressed as a quadratic form $(\mathbf{Y}'\mathbf{A}\mathbf{Y})$, where the \mathbf{A} matrices are $(\mathbf{I} - \mathbf{H})$, $(\mathbf{H} - \frac{1}{n}\mathbf{J})$, and $(\mathbf{I} - \frac{1}{n}\mathbf{J})$.

Knowing that the sums of squares are quadratic forms allows the statistician to prove various important results. Figure 12.3 provides a graphical representation of decomposition of the total deviation (as in (12.35)) found in ANOVA.

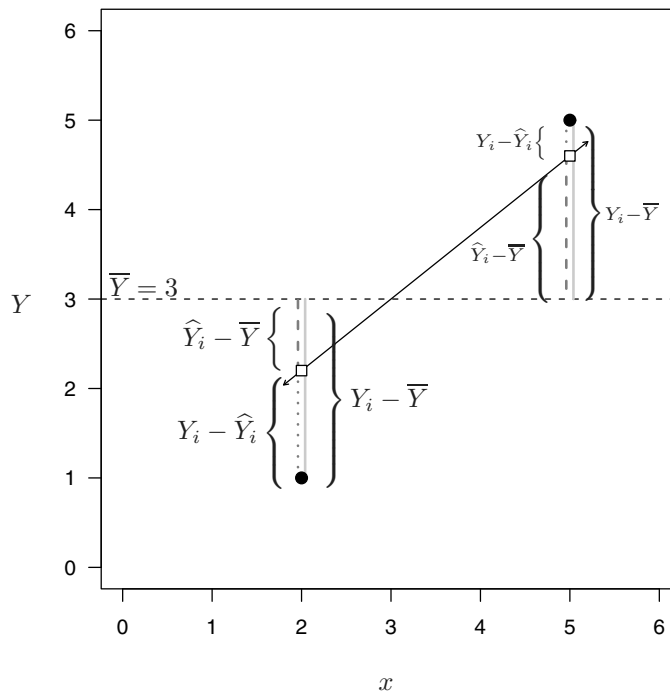


FIGURE 12.3: Graphical representation of the sum of squares partition

12.9.3 Coefficient of Determination

Figure 12.4 shows three different scatterplots of bivariate data. The points in the first scatterplot fall exactly on a straight line. Consequently, 100% of the variability in the y values can be attributed to the linear relationship between y and x . The points in the second scatterplot do not fall exactly along a line; however, the deviations from the least squares line ($Y_i - \hat{Y}_i$) compared to the total deviations ($Y_i - \bar{Y}$) are relatively small. This makes it reasonable to conclude that a large proportion of the variability in y can be attributed to the linear relationship between y and x . The third scatterplot shows both large deviations from the least squares line as well as large total deviations. In this case, a linear relationship between y and x is not overly helpful in explaining the variability of the y_i s exhibited in the scatterplot.

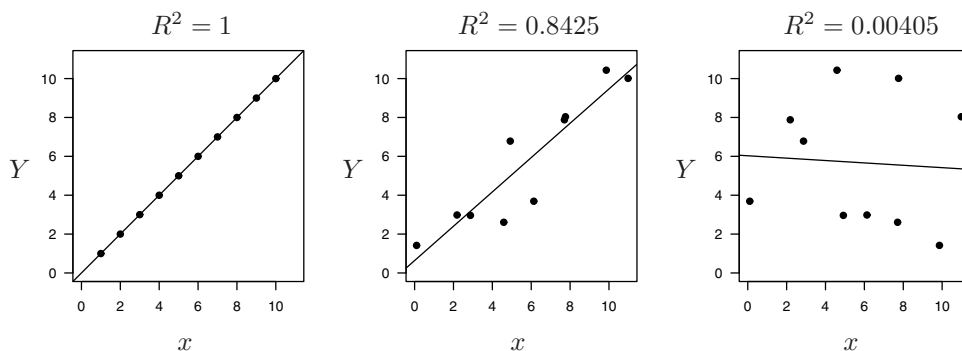


FIGURE 12.4: Scatterplots to illustrate values of R^2

The sum of squares due to error (SSE) can be interpreted as the amount of variability in Y that is unexplained by a linear model. Since SSE ($\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$) is smaller than the sum of squared deviations of any other line, $SSE \leq SST$. Note that only in the case of a horizontal line would $SSE = SST$. Consequently, the ratio $\frac{SSE}{SST}$ represents the proportion of variability that cannot be explained by the linear regression model. In an analogous fashion, R^2 , the **coefficient of determination**, represents the proportion of variability in the Y_i s that can be explained by the simple linear regression model where

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST} \quad (12.43)$$

When working with linear regression models with $p - 1$ explanatory variables, R^2 is interpreted as the proportion of variability in the Y_i s that can be explained with a linear model containing the variables x_1, x_2, \dots, x_{p-1} . Since adding more x -variables to the regression model can only increase R^2 (as SSE never increases as more variables are added to a model and SST is constant for any set of Y_i values), a measure is needed that takes into account how many variables are in a model to determine the most appropriate variables to include. Such a measure is the **adjusted coefficient of determination**, R_a^2 . R_a^2 is

computed by dividing each sum of squares by its associated degrees of freedom. That is,

$$R_a^2 = 1 - \frac{\frac{SSE}{n-p}}{\frac{SST}{n-1}} = 1 - \frac{MSE}{MST} \quad (12.44)$$

Although R^2 and R_a^2 provide a certain measure of “goodness-of-fit” for the fitted model, they should be used with caution and never as the sole criterion for determining which among several models is best.

12.9.4 Extra Sum of Squares

An extra sum of squares measures the marginal increase in the regression sum of squares when one or more variables are added to a regression model. The marginal increase when adding x_2 to a model that already contains x_1 will be denoted as

$$SSR(x_2|x_1) = SSR(x_2, x_1) - SSR(x_1) \quad (12.45)$$

which is equivalent to

$$SSR(x_2|x_1) = SSE(x_1) - SSE(x_1, x_2).$$

When the regression model contains r x -variables, there are $r!$ possible decompositions of the x -variables.

Example 12.9 Consider the case where $r = 3$. What are the six decompositions of $SSR(x_1, x_2, x_3)$?

Solution:

$$\begin{aligned} SSR(x_1, x_2, x_3) &= SSR(x_1) + SSR(x_2|x_1) + SSR(x_3|x_1, x_2) \\ SSR(x_1, x_3, x_2) &= SSR(x_1) + SSR(x_3|x_1) + SSR(x_2|x_1, x_3) \\ SSR(x_2, x_1, x_3) &= SSR(x_2) + SSR(x_1|x_2) + SSR(x_3|x_1, x_2) \\ SSR(x_2, x_3, x_1) &= SSR(x_2) + SSR(x_3|x_2) + SSR(x_1|x_2, x_3) \\ SSR(x_3, x_1, x_2) &= SSR(x_3) + SSR(x_1|x_3) + SSR(x_2|x_1, x_3) \\ SSR(x_3, x_2, x_1) &= SSR(x_3) + SSR(x_2|x_3) + SSR(x_1|x_2, x_3) \quad \blacksquare \end{aligned} \quad (12.47)$$

Example 12.10 The data frame `HSwrestler` contains information on nine variables for a group of 78 high school wrestlers that was collected by the human performance lab at Appalachian State University. The variables are `AGE` (in years), `HT` (height in inches), `WT` (weight in pounds), `ABS` (abdominal skinfold measure), `TRICEPS` (tricep skinfold measure), `SUBSCAP` (subscapular skinfold measure), `HWFAT` (hydrostatic determination of fat), `TANFAT` (Tanita determination of fat), and `SKFAT` (skinfold determination of fat). Use `S` to obtain the ANOVA results when hydrostatic fat (Y) is regressed on `ABS` (x_1), `TRICEPS` (x_2), and `SUBSCAP` (x_3) to verify empirically the results from (12.46) and (12.47).

Solution: The order variables specified in `S` impact the ANOVA table since the sums of squares reported are conditional sums of squares. First, the $SSR(x_1, x_2, x_3)$ is computed using the formula from Table 12.3 on page 584:

```
> attach(HSwrestler)
> Y <- HWFAT
> x1 <- ABS
```

```

> x2 <- TRICEPS
> x3 <- SUBSCAP
> n <- 78
> X <- cbind(rep(1, n), x1, x2, x3)
> Y <- matrix(HWFAT, nrow=n)
> J <- matrix(rep(1, n*n), nrow=n)
> H <- X%*%solve(t(X)%*%X)%*%t(X)
> b <- solve(t(X)%*%X)%*%t(X)%*%Y
> SSR <- t(b)%*%t(X)%*%Y-(1/n)*t(Y)%*%J%*%Y
> SSE <- t(Y)%*%Y - t(b)%*%t(X)%*%Y
> SST0 <- t(Y)%*%Y -(1/n)*t(Y)%*%J%*%Y
> SS <- rbind(SSR, SSE, SST0)
> r.names <- c("SSR", "SSE", "SST0")
> c.names <- c("Sum of Squares")
> dimnames(SS) <- list(r.names, c.names)
> SS
      Sum of Squares
SSR      5317.252
SSE       700.540
SST0     6017.792

```

Computing SSR , SSE , and SST with (12.42), (12.40), and (12.41), respectively, yields

```

> ID <- diag(nrow=n) # n*n Identity matrix
> SSRa <- t(Y)%*%(H-(1/n)*J)%*%Y
> SSEa <- t(Y)%*%(ID-H)%*%Y
> SST0a <- t(Y)%*%(ID-(1/n)*J)%*%Y
> SSa <- rbind(SSRa, SSEa, SST0a)
> r.names <- c("SSR", "SSE", "SST0")
> c.names <- c("Sum of Squares")
> dimnames(SSa) <- list(r.names, c.names)
> SSa
      Sum of Squares
SSR      5317.252
SSE       700.540
SST0     6017.792

```

Note that the order of the x_i s does not impact the computation of SSR :

```

> mod123 <- lm(Y~x1+x2+x3)
> anova(mod123)

```

Table 12.4: ANOVA table for $\text{mod123} <- \text{lm}(Y \sim x_1 + x_2 + x_3)$

anova	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	5072.8	5072.8	535.858	$< 2.2e - 16$
x2	1	242.2	242.2	25.581	$2.984e - 06$
x3	1	2.2	2.2	0.237	0.6278
Residuals	74	700.5	9.5		

$$\begin{aligned}
 SSR(x_1, x_2, x_3) &= SSR(x_1) + SSR(x_2|x_1) + SSR(x_3|x_1, x_2) \\
 5317.2 &\stackrel{?}{=} 5072.8 + 242.2 + 2.2 \\
 5317.2 &= 5317.2
 \end{aligned}$$

```

> mod321 <- lm(Y~x3+x2+x1)
> anova(mod321)
> detach(HSwrestler)

```

Table 12.5: ANOVA table for `mod321 <- lm(Y~x3+x2+x1)`

anova	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x3	1	4939.0	4939.0	521.720	$< 2.2e - 16$
x2	1	204.6	204.6	21.616	$1.422e - 05$
x1	1	173.6	173.6	18.341	$5.473e - 05$
Residuals	74	700.5	9.5		

$$\begin{aligned}
 SSR(x_3, x_2, x_1) &= SSR(x_3) + SSR(x_2|x_3) + SSR(x_1|x_2, x_3) \\
 5317.2 &\stackrel{?}{=} 4939.0 + 204.6 + 173.6 \\
 5317.2 &= 5317.2
 \end{aligned}$$



12.9.4.1 Tests on a Single Parameter

To test whether the term $\beta_k x_k$ can be dropped from a multiple regression model, the hypotheses of interest are

$$H_0 : \beta_k = 0 \quad \text{versus} \quad H_1 : \beta_k \neq 0$$

It was shown earlier in (12.33) that $t_{\text{obs}} = \hat{\beta}_k / s_{\hat{\beta}_k}$ could be used as an appropriate test statistic. It is also possible to test $\beta_k = 0$ using a general linear test statistic that involves extra sum of squares. Consider a regression model with three predictor variables (which represent the full model). To test the hypothesis

$$H_0 : \beta_2 = 0 \quad \text{versus} \quad H_1 : \beta_2 \neq 0,$$

a reduced model where $\beta_2 x_2$ has been eliminated from the full model is computed. The general linear test statistic is

$$F_{\text{obs}} = \frac{\frac{SSR(F) - SSR(R)}{df_F - df_R}}{\frac{SSE}{df}} \quad (12.48)$$

where F stands for the full model and R stands for the reduced model. SSE is the sum of squares error for the full model and df is the degrees of freedom for error for the full model. This F_{obs} follows an F distribution with $(df_F - df_R, df)$ degrees of freedom, under the assumptions that H_0 is true and the normal error linear model assumptions are satisfied.

Example 12.11 Use the data frame `HSwrestler` to show the equivalence between the t_{obs} and F_{obs} values when testing the hypothesis $H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 \neq 0$ when regressing `HWFAT` (Y) on `ABS` (x_1), `TRICEPS` (x_2), and `SUBSCAP` (x_3).

Solution: The important concept to remember is that both tests assume x_1 and x_3 are in the model. Consequently, x_2 must be entered into the model last.

```
> attach(HSwrestler)
> Y <- HWFAT
> x1 <- ABS
> x2 <- TRICEPS
> x3 <- SUBSCAP
> mod132 <- lm(Y~x1+x3+x2)
> anova(mod132)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1  5072.8   5072.8  535.858 < 2.2e-16 ***
x3         1   132.6    132.6   14.005 0.0003577 ***
x2         1   111.8    111.8   11.814 0.0009682 ***
Residuals 74   700.5     9.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(mod132)[3,4]          # Fobs value
[1] 11.81363
> summary(mod132)

Call:
lm(formula = Y ~ x1 + x3 + x2)

Residuals:
    Min       1Q   Median       3Q      Max
-6.4316 -2.4258 -0.4800  2.2797  9.5509

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.06997    0.65592   3.156 0.002315 **
x1           0.31894    0.07447   4.283 5.47e-05 ***
x3           0.06632    0.13622   0.487 0.627819
x2           0.46069    0.13404   3.437 0.000968 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.077 on 74 degrees of freedom
Multiple R-Squared:  0.8836,    Adjusted R-squared:  0.8789
F-statistic: 187.2 on 3 and 74 DF,  p-value: < 2.2e-16

> summary(mod132)$coefficients[4,3]^2 # tobs value squared
[1] 11.81363
> detach(HSwrestler)
```

The value of F_{obs} may also be found with `drop1(mod132, test="F")`.

From the anova output (12.49), calculate

$$\begin{aligned} SSR(F) &= 5072.8 + 132.6 + 111.8 = 5317.2 \\ SSR(R) &= 5072.8 + 132.6 = 5025.4 \end{aligned}$$

which gives an F_{obs} value of

$$F_{\text{obs}} = \frac{\frac{SSR(F) - SSR(R)}{df_F - df_R}}{\frac{SSE}{df}} = \frac{\frac{5317.2 - 5025.4}{3 - 2}}{\frac{700.5}{74}} = 11.81.$$

The t_{obs} value is

$$t_{\text{obs}} = \frac{\hat{\beta}_2}{s\hat{\beta}_2} = \frac{.46069}{.13404} = 3.437.$$

Recall from (6.29) that $t_{1-\alpha/2;\nu}^2 = f_{1-\alpha;1,\nu}$. So, the equivalence between the t_{obs} and F_{obs} values is equivalent to showing that $(3.437)^2 = 11.81$, which it does to two decimal places. ■

12.9.4.2 Tests on Subsets of the Regression Parameters

Traditional tests of hypotheses on individual regression coefficients generated with the `S` command `anova(lm.object)` are partitions of the regression sum of squares. Frequently, the user may want to test hypotheses containing a subset of the regression parameters. To test whether a subset of the regression parameters are equal to zero requires the general linear statistic in (12.48). Consider a full model with three predictors, x_1 , x_2 , and x_3 . To see if $\beta_1 x_{i1}$ and $\beta_2 x_{i2}$ can be dropped from the model, the test of hypothesis is

$$H_0 : \beta_1 = \beta_2 = 0 \text{ versus } H_1 : \text{at least one } \beta_i \neq 0 \text{ for } i = 1, 2$$

The full model is $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$ while the reduced model is $Y_i = \beta_0 + \beta_3 x_{i3} + \varepsilon_i$. Consequently, the general linear test statistic will be

$$\begin{aligned} F_{\text{obs}} &= \frac{\frac{SSR(F) - SSR(R)}{df_F - df_R}}{\frac{SSE}{df}} = \frac{\frac{SSR(x_1, x_2, x_3) - SSR(x_3)}{3 - 1}}{\frac{SSE}{n - p}} \\ &= \frac{\frac{SSR(x_1, x_2|x_3)}{2}}{MSE} = \frac{MSR(x_1, x_2|x_3)}{MSE} \end{aligned}$$

Example 12.12 Use the data frame `HSwrestler` and test whether $\beta_1 x_{i1}$ and $\beta_2 x_{i2}$ can be dropped from the full model $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$, where $x_1 = \text{ABS}$, $x_2 = \text{TRICEPS}$, and $x_3 = \text{SUBSCAP}$ using the general linear test approach with a 0.01 significance level.

Solution: To test $H_0 : \beta_1 = \beta_2 = 0$ versus $H_1 : \text{at least one } \beta_i \neq 0 \text{ for } i = 1, 2$, one must ensure x_3 is the first variable specified in the model to facilitate the solution:

```
> attach(HSwrestler)
> Y <- HWFAT
> x1 <- ABS
```

```
> x2 <- TRICEPS
> x3 <- SUBSCAP
> mod312 <- lm(Y~x3+x1+x2)
> anova(mod312)
```

Table 12.6: ANOVA table for Example 12.12 on the preceding page

anova	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x3	1	4939.0	4939.0	521.720	$< 2.2e - 16$
x1	1	266.4	266.4	28.143	$1.129e - 06$
x2	1	111.8	111.8	11.814	0.0009682
Residuals	74	700.5	9.5		

$$F_{\text{obs}} = \frac{\frac{SSR(F) - SSR(R)}{df_F - df_R}}{\frac{SSE}{df}}$$

$$= \frac{(4939 + 266.4 + 111.8) - (4939)}{\frac{3 - 1}{\frac{700.5}{74}}} = 19.98$$

```
> pf(19.98, 2, 74, lower.tail=FALSE) # Only in R
[1] 1.152745e-07
```

Since p -value = $\mathbb{P}(F_{2,74} \geq 19.98) = 1.15 \times 10^{-7} < 0.01$, reject H_0 and declare the results statistically significant. The evidence suggests that ABS and TRICEPS should not be dropped from a model that already contains SUBSCAP.

Another approach to test $H_0 : \beta_1 = \beta_2 = 0$ is to use `anova()` on the reduced and full models:

```
> mod3 <- lm(Y~x3)
> anova(mod3, mod312)
Analysis of Variance Table

Model 1: Y ~ x3
Model 2: Y ~ x3 + x1 + x2
  Res.Df    RSS Df Sum of Sq    F    Pr(>F)
1     76 1078.80
2     74  700.54  2    378.26 19.978 1.154e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> detach(HSwrestler)
```



12.10 General Linear Hypothesis

Tests on individual parameters and on subsets of parameters can be expressed in a much more general fashion that provides fantastic flexibility in testing. The **general linear hypotheses** are

$$H_0 : \mathbf{K}\boldsymbol{\beta} = \mathbf{m} \quad \text{versus} \quad H_1 : \mathbf{K}\boldsymbol{\beta} \neq \mathbf{m} \quad (12.50)$$

where \mathbf{K} is a $q \times p$ matrix of rank $q \leq p$ with each row corresponding to one partial hypothesis and \mathbf{m} is a numerical vector. When working with the normal error model, where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ and $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1})$, the parameter vector $\mathbf{K}\boldsymbol{\beta}$ is estimated by $\mathbf{K}\hat{\boldsymbol{\beta}}$, which is a linear combination of normally distributed random variables. This implies $\mathbf{K}\hat{\boldsymbol{\beta}} \sim N(\mathbf{K}\boldsymbol{\beta}, \sigma^2 \mathbf{K}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}')$. When \mathbf{K} is a vector ($q = 1$), the null hypothesis $\mathbf{K}\boldsymbol{\beta} = \mathbf{m}_0$ is tested with a t -statistic since

$$t = \frac{\mathbf{K}\hat{\boldsymbol{\beta}} - \mathbf{m}_0}{\sqrt{\hat{\sigma}^2 \mathbf{K}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}'}} \sim t_{n-p, \gamma} \quad (12.51)$$

where

$$\gamma = \frac{\mathbf{K}\boldsymbol{\beta} - \mathbf{m}_0}{\sqrt{\sigma^2 \mathbf{K}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}'}} \quad (12.52)$$

Under the null hypothesis, $\gamma = 0$ and (12.51) is a central t -distribution with $n - p$ degrees of freedom. In the general linear hypothesis, only a two-sided alternative is given; however, when $q = 1$, the one-sided alternative, $H_1 : \mathbf{K}\boldsymbol{\beta} > \mathbf{m}$ or $H_1 : \mathbf{K}\boldsymbol{\beta} < \mathbf{m}$, may be specified and tested using (12.51).

When the rank of \mathbf{K} is greater than one ($q > 1$), the quantity

$$\frac{(\mathbf{K}\hat{\boldsymbol{\beta}} - \mathbf{m})'(\mathbf{K}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}')^{-1}(\mathbf{K}\hat{\boldsymbol{\beta}} - \mathbf{m})}{q\hat{\sigma}^2} \sim F_{q, n-p, \lambda} \quad (12.53)$$

where

$$\lambda = \frac{1}{\sigma^2}(\mathbf{K}\boldsymbol{\beta} - \mathbf{m})'(\mathbf{K}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}')^{-1}(\mathbf{K}\boldsymbol{\beta} - \mathbf{m}) \quad (12.54)$$

is used to test the null hypothesis $H_0 : \mathbf{K}\boldsymbol{\beta} = \mathbf{m}$. For details, see Graybill (1976). Note that Graybill does offer a slightly different definition of λ (his λ is that of (12.54) divided by 2), but this does not complicate the exposition. This text will use the definition used by S and Rao (1973). By using (12.54), the relationship between γ and λ is $\lambda = \gamma^2$. The square of the non-central t -statistic with non-centrality parameter γ is distributed as a non-central F -statistic with non-centrality parameter $\lambda = \gamma^2$. In other words, $t_{\nu, \gamma}^2 = f_{1, \nu, \gamma^2}$. The power of the test $H_0 : \mathbf{K}\boldsymbol{\beta} = \mathbf{m}$ as a function of λ for a given α is $\text{Power}(\lambda) = \mathbb{P}(F_{\nu_1, \nu_2, \lambda} > f_{1-\alpha; \nu_1, \nu_2})$.

When $H_0 : \mathbf{K}\boldsymbol{\beta} = \mathbf{m}$ is true,

$$\frac{(\mathbf{K}\hat{\boldsymbol{\beta}} - \mathbf{m})'(\mathbf{K}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}')^{-1}(\mathbf{K}\hat{\boldsymbol{\beta}} - \mathbf{m})}{\sigma^2} \sim \chi_q^2 \quad (12.55)$$

and

$$F_{\text{obs}} = \frac{(\mathbf{K}\hat{\boldsymbol{\beta}} - \mathbf{m})'(\mathbf{K}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}')^{-1}(\mathbf{K}\hat{\boldsymbol{\beta}} - \mathbf{m})}{q\hat{\sigma}^2} \sim F_{q, n-p} \quad (12.56)$$

To test $H_0 : \mathbf{K}\boldsymbol{\beta} = \mathbf{m}$ at the α significance level, reject H_0 and conclude H_1 when $\mathbb{P}(F_{q, n-p} \geq F_{\text{obs}}) < \alpha$. The function `glht()` in the R package `multcomp` can greatly ease the computation of general linear hypotheses, especially when testing for linear relationships among the β_j s.

Example 12.13 ▷ *General Linear Model* ◁ Use a general linear hypothesis with $\alpha = 0.05$ to

- (a) Test whether $\beta_2 x_{i2}$ and $\beta_3 x_{i3}$ can be dropped from the model $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$, where $x_1 = \text{ABS}$, $x_2 = \text{TRICEPS}$, and $x_3 = \text{SUBSCAP}$, using information from the data frame `HSwrestler`.
- (b) Test the two linear relationships

$$\begin{aligned} 2\beta_1 + \beta_2 &= \beta_3 \\ -5\beta_1 + \beta_3 &= 0.20 \end{aligned}$$

- (c) Test whether $\beta_2 = \beta_3$.

Solution: The answers are as follows:

- (a)

Step 1: **Hypotheses** —

$$H_0 : \mathbf{K}\boldsymbol{\beta} = \mathbf{m} \quad \text{versus} \quad H_1 : \mathbf{K}\boldsymbol{\beta} \neq \mathbf{m}$$

$$\text{where } \mathbf{K} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \text{ and } \mathbf{m} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Step 2: **Test Statistic** — Under the assumption H_0 is true,

$$F_{\text{obs}} = \frac{(\mathbf{K}\hat{\boldsymbol{\beta}} - \mathbf{m})'(\mathbf{K}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}')^{-1}(\mathbf{K}\hat{\boldsymbol{\beta}} - \mathbf{m})}{q\hat{\sigma}^2} \sim F_{q, n-p}.$$

Step 3: **Rejection Region Calculations** — Because $F_{\text{obs}} \sim F_{2, 74}$ and this is a one-tailed test, the rejection region is $F_{\text{obs}} > f_{0.95; 2, 74} = 3.12$. The value of the standardized test statistic is $F_{\text{obs}} = 19.98$:

$$F_{\text{obs}} = \frac{\begin{bmatrix} 0.319 & 0.461 \end{bmatrix} \begin{bmatrix} 1819.5992 & 251.6221 \\ 251.6221 & 561.7319 \end{bmatrix} \begin{bmatrix} 0.319 \\ 0.461 \end{bmatrix}}{2(9.47)} = 19.98$$

Step 4: **Statistical Conclusion** — The φ -value is $\mathbb{P}(F_{2, 74} \geq 19.98) = 0+$, where $0+$ means that the φ -value is zero to more than six decimal places without being equal to zero.

- I. From the rejection region, reject H_0 because 19.98 is greater than 3.12.
- II. From the φ -value, reject H_0 because the φ -value = $0+$ is less than 0.05.

Step 5: **English Conclusion** — There is strong evidence to suggest a linear relationship exists between ABS, TRICEPS, and HWFAT, suggesting neither variable should be dropped from a model that currently contains SUBSCAP.

```
> attach(HSwrestler)
> Y <- HWFAT
> x1 <- ABS
> x2 <- TRICEPS
```

```

> x3 <- SUBSCAP
> mod312 <- lm(Y~x3+x1+x2)
> K <- matrix(c(0,0,1,0,0,0,0,1), byrow=TRUE, nrow=2)
> Q <- qr(K)$rank
> b <- matrix(coef(mod312), ncol=1)
> m <- matrix(c(0,0), byrow=TRUE, nrow=2)
> XTXI <- summary(mod312)$cov.unscaled
> NUM <- t(K%*%b-m)%*%solve(K%*%XTXI%*%t(K))%*%(K%*%b-m)
> MSE <- anova(mod312)[4,3]
> Fobs <- NUM/(Q*MSE)
> c(Fobs,1-pf(Fobs, Q,74))
[1] 1.997828e+01 1.154032e-07

```

Using `glht()` from the `multcomp` package,

```

> library(multcomp)
Loading required package: mvtnorm
> summary(glht(mod312, linct=K, rhs=c(0,0)), test=Ftest())

```

General Linear Hypotheses

Linear Hypotheses:

```

      Estimate
1 == 0  0.3189
2 == 0  0.4607

```

Global Test:

```

      F DF1 DF2      Pr(>F)
1 19.98   2   74 1.154e-07

```

(b)

Step 1: **Hypotheses** —

$$H_0 : \mathbf{K}\boldsymbol{\beta} = \mathbf{m} \quad \text{versus} \quad H_1 : \mathbf{K}\boldsymbol{\beta} \neq \mathbf{m}$$

$$\text{where } \mathbf{K} = \begin{bmatrix} 0 & 2 & 1 & -1 \\ 0 & -5 & 0 & 1 \end{bmatrix} \text{ and } \mathbf{m} = \begin{bmatrix} 0 \\ .2 \end{bmatrix}.$$

Step 2: **Test Statistic** — Under the assumption H_0 is true,

$$F_{\text{obs}} = \frac{(\mathbf{K}\hat{\boldsymbol{\beta}} - \mathbf{m})'(\mathbf{K}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}')^{-1}(\mathbf{K}\hat{\boldsymbol{\beta}} - \mathbf{m})}{q\hat{\sigma}^2} \sim F_{q, n-p}.$$

Step 3: **Rejection Region Calculations** — Because $F_{\text{obs}} \sim F_{2, 74}$ and this is a one-tailed test, the rejection region is $F_{\text{obs}} > f_{0.95; 2, 74} = 3.12$. The value of the standardized test statistic is $F_{\text{obs}} = 0.062$:

$$F_{\text{obs}} = \frac{\begin{bmatrix} -0.0091 & -0.0709 \end{bmatrix} \begin{bmatrix} 676.0834 & 297.8446 \\ 297.8446 & 146.8894 \end{bmatrix} \begin{bmatrix} -0.0091 \\ -0.0709 \end{bmatrix}}{2(9.47)} = 0.0623$$

Step 4: **Statistical Conclusion** — The φ -value is $\mathbb{P}(F_{2,74} \geq 0.0623) = 0.94$.

- I. From the rejection region, fail to reject H_0 because 0.062 is less than 3.12.
- II. From the φ -value, fail to reject H_0 because the φ -value = 0.94 is greater than 0.05.

Step 5: **English Conclusion** — There is no evidence to suggest the postulated relationships $\mathbf{K}\boldsymbol{\beta} \neq \mathbf{m}$.

```
> K <- matrix(c(0,2,1,-1,0,-5,0,1), byrow=TRUE, nrow=2)
> summary(glht(mod312, linfct=K, rhs=c(0,0.2)), test=Ftest())
```

General Linear Hypotheses

Linear Hypotheses:

```
      Estimate
1 == 0   -0.00912
2 == 0.2  0.12911
```

Global Test:

```
      F DF1 DF2 Pr(>F)
1 0.0623  2  74 0.9396
```

(c)

Step 1: **Hypotheses** —

$$H_0 : \mathbf{K}\boldsymbol{\beta} = \mathbf{m} \quad \text{versus} \quad H_1 : \mathbf{K}\boldsymbol{\beta} \neq \mathbf{m}$$

where $\mathbf{K} = \begin{bmatrix} 0 & 0 & 1 & -1 \end{bmatrix}$ and $\mathbf{m} = \begin{bmatrix} 0 \end{bmatrix}$.

Step 2: **Test Statistic** — Under the assumption H_0 is true,

$$F_{\text{obs}} = \frac{(\mathbf{K}\hat{\boldsymbol{\beta}} - \mathbf{m})'(\mathbf{K}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}')^{-1}(\mathbf{K}\hat{\boldsymbol{\beta}} - \mathbf{m})}{q\hat{\sigma}^2} \sim F_{q, n-p}.$$

Step 3: **Rejection Region Calculations** — Because $F_{\text{obs}} \sim F_{1,74}$ and this is a one-tailed test, the rejection region is $F_{\text{obs}} > f_{0.95;1,74} = 3.97$. The value of the standardized test statistic is $F_{\text{obs}} = 0.062$:

$$F_{\text{obs}} = \frac{[-0.1418][332.3932][-0.1418]}{2(9.47)} = 0.7056.$$

Step 4: **Statistical Conclusion** — The φ -value is $\mathbb{P}(F_{1,74} \geq 0.7056) = 0.40$.

- I. From the rejection region, fail to reject H_0 because 0.7056 is less than 3.97.
- II. From the φ -value, fail to reject H_0 because the φ -value = 0.40 is greater than 0.05.

Step 5: **English Conclusion** — There is no evidence to suggest $\beta_2 \neq \beta_3$.

```
> K <- matrix(c(0,0,1,-1), byrow=TRUE, nrow=1)
> summary(glht(mod312, linct=K, rhs=0), test=Ftest())
```

General Linear Hypotheses

Linear Hypotheses:

```
Estimate
== 0 -0.1418
```

Global Test:

```
F DF1 DF2 Pr(>F)
1 0.7056 1 74 0.4036
> detach(HSwrestler)
```



12.11 Model Selection and Validation

The process of selecting a model either for predictive or explanatory purposes involves several procedures, where the order of the procedures is not always the same. One needs always to bear in mind that regression analysis is simply a tool to understand the structure of data. In what follows, general methods to select models, to verify assumptions, and to perform transformations on both the response and predictor variables are discussed. Although an order is presented for building a model, the analyst should always be alert for an unexpected structure in the data and be flexible in his assessment of the model.

12.11.1 Testing-Based Procedures

When building a model, it is desirable to select the “best” subset of predictors that explains the data in the simplest fashion. Adding too many variables wastes degrees of freedom and adds unwanted noise to the problem, increases the risk of adding variables that measure the same quantity, as well as increases the effort needed to measure the redundant predictors. There are two basic approaches one can take to select variables: 1) a stepwise testing strategy that compares successive models and 2) a criterion approach that attempts to maximize some measure of goodness-of-fit.

12.11.1.1 Backward Elimination

Backward elimination begins with a model containing all potential x -variables and identifies the one with the largest φ -value. This can be done by looking at the φ -values for the t -values of the $\hat{\beta}_i, i = 1, \dots, p-1$ using the function `summary()` or using the φ -values from the R function `drop1()`. If the variable with the largest φ -value is above a predetermined value, α_{crit} , that x -variable is dropped. A model with the remaining x -variables is then fit and the procedure continues until all the φ -values for the remaining variables in the model are below the predetermined α_{crit} . The α_{crit} is sometimes referred to as the “ φ -to-remove” and is typically set to 15 or 20%.

12.11.1.2 Forward Selection

Forward selection starts with no variables in the model and then adds the x -variable that produces the smallest φ -value below α_{crit} when included in the model. This procedure is continued until no new predictors can be added. The user can determine the variable

that produces the smallest φ -value by regressing the response variable on the x_i s one at a time using `lm()` and `summary()` or by using the function `add1()`.

12.11.1.3 Stepwise Regression

This is a combination of backward elimination and forward selection. This technique allows variables that were either removed or added early in the procedure to reenter or exit the model later in the process. At each stage, a variable may be added or removed.

Testing-based procedures are relatively straightforward to implement; however, they do have some drawbacks. One of the chief weaknesses of testing-based procedures is ending up with a model that is overly parsimonious. When the analyst has a firm grasp of the subject matter, the analyst may want to include predictors that appear to have no statistical significance. Although predictors can be added to a model developed from a testing-based perspective, the idea of adding predictors that are not necessarily significant conforms more to a criterion-based procedure.

12.11.1.4 Criterion-Based Procedures

There are several well-defined optimality criteria used in model building including R_a^2 (R^2 adjusted), Mallows's C_p , Bayes Information Criterion (BIC), and Akaike Information Criterion (AIC). R_a^2 is used instead of R^2 since R^2 will always increase with the addition of more variables to the model. Recall that $R_a^2 = 1 - ((n-1)/(n-p)) \cdot (SSE/SST)$.

The C_p statistic is defined as $C_p = SSE/\hat{\sigma}^2 + 2p - n$, where $\hat{\sigma}^2$ is from the model with all predictors and SSE is for the model with p parameters. When all p parameters are used in the model, $C_p = p$. A model with a bad fit will produce a C_p much bigger than p . Desirable models have small p and C_p less than or equal to p . It is common practice to plot C_p against p .

Recall that $\ln \mathcal{L}(\beta, \sigma^2 | \mathbf{X})$ is called the log-likelihood function. The BIC for linear regression models is defined as $-2 \max(\ln \mathcal{L}(\beta, \sigma^2 | \mathbf{X})) + p \cdot \ln(n) = n \ln(SSE/n) + p \cdot \ln(n) + \text{constant}$, while the AIC for linear regression models is defined as $-2 \max(\ln \mathcal{L}(\beta, \sigma^2 | \mathbf{X})) + 2p = n \ln(SSE/n) + 2p + \text{constant}$. Since the constant is the same for a given data set and error distribution, it can be ignored when comparing models based on the same data. This is what the function `stepAIC()` does.

There are various S functions including `step()` to perform criterion-based searches. Unfortunately, the function `step()` in R is not equivalent to the function `step()` in S-PLUS. Fortunately, the function `stepAIC()` in the MASS package is an equivalent function in both R and S-PLUS. The goal when using BIC or AIC is to create a model that minimizes either BIC or AIC. Both AIC and BIC search for models that have small SSE . However, BIC penalizes larger models more so than does AIC (assuming $n > e^2 = 7.39$). Consequently, BIC will favor smaller models than will AIC. When building a model to be used for predictive purposes, AIC will generally be favored over BIC. For those using R, the package `leaps` contains the function `regsubsets()`, which is very useful for computing R_a^2 and Mallows's C_p .

Example 12.14 ▷ *Model Selection with HS`wrestler`* ◁ Create a model for predicting wrestlers' hydrostatic fat (HWFAT) for the data frame `HSwrestler`.

- (a) Use backward elimination with the predictors AGE, HT, WT, ABS, TRICEPS, and SUBSCAP and an α_{crit} of 0.20.
- (b) Use forward selection with an α_{crit} of 0.20.

- (c) Use the function `regsubsets` in the R package `leaps` to select a model using R_a^2 as the criterion.
- (d) Use the function `regsubsets` in the R package `leaps` to select a model using Mallows's C_p as the criterion.
- (e) Use AIC as the criterion for selecting a model.
- (f) Use BIC as the criterion for selecting a model.

Solution: All output is from R.

- (a) Backward elimination starts with all the variables in the model and eliminates variables with the largest (least significant) ϕ -values:

```
> attach(HSwrestler)
> # Backward elimination showing all steps
> reg.all <- lm(HWFAT ~ AGE + HT + WT + ABS + TRICEPS + SUBSCAP)
> summary(reg.all)$coefficients
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	13.29369860	9.63026704	1.3804081	1.717917e-01
AGE	-0.32893403	0.32157778	-1.0228755	3.098393e-01
HT	-0.06730905	0.16050751	-0.4193514	6.762255e-01
WT	-0.01365183	0.02590783	-0.5269385	5.998789e-01
ABS	0.37141976	0.08836595	4.2032001	7.548985e-05
TRICEPS	0.38742647	0.13761017	2.8153912	6.301113e-03
SUBSCAP	0.11405213	0.14192779	0.8035927	4.243145e-01

Note that HT has the largest ϕ -value of $6.762255e-01$, so it is eliminated from the model:

```
> reg.m1 <- lm(HWFAT ~ AGE + WT + ABS + TRICEPS + SUBSCAP)
> summary(reg.m1)$coefficients
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	9.69230054	4.33250597	2.2371119	2.837559e-02
AGE	-0.33352357	0.31954680	-1.0437393	3.000978e-01
WT	-0.02084061	0.01931391	-1.0790465	2.841686e-01
ABS	0.38259027	0.08377184	4.5670510	1.996022e-05
TRICEPS	0.39737898	0.13477014	2.9485685	4.302189e-03
SUBSCAP	0.11175170	0.14100772	0.7925218	4.306601e-01

Note that SUBSCAP has the largest ϕ -value of $4.306601e-01$, so it is eliminated from the model:

```
> reg.m2 <- lm(HWFAT ~ AGE + WT + ABS + TRICEPS)
> summary(reg.m2)$coefficients
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	9.94025577	4.31017288	2.3062313	2.393916e-02
AGE	-0.38382444	0.31238134	-1.2287048	2.231289e-01
WT	-0.01585418	0.01821376	-0.8704507	3.869075e-01
ABS	0.39968360	0.08074124	4.9501789	4.621329e-06
TRICEPS	0.46942072	0.09924414	4.7299591	1.068468e-05

Note that WT has the largest ϕ -value of 3.869075e-01, so it is eliminated from the model:

```
> reg.m3 <- lm(HWFAT ~ AGE + ABS + TRICEPS)
> summary(reg.m3)
```

Call:

```
lm(formula = HWFAT ~ AGE + ABS + TRICEPS)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-5.8374 -2.0468 -0.4215  2.3076  7.9850
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.61606    4.23272   2.508  0.0143 *
AGE          -0.53309    0.26067  -2.045  0.0444 *
ABS           0.35643    0.06354   5.610 3.32e-07 ***
TRICEPS       0.46561    0.09898   4.704 1.16e-05 ***
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 2.998 on 74 degrees of freedom

Multiple R-Squared: 0.8895, Adjusted R-squared: 0.885

F-statistic: 198.5 on 3 and 74 DF, p-value: < 2.2e-16

The remaining ϕ -values for AGE, ABS, and TRICEPS are all less than 0.20, so the model is composed of these three variables based on backward elimination.

Alternately, the function `drop1()` can be used in R:

```
> drop1(lm(HWFAT ~ AGE + HT + WT + ABS + TRICEPS + SUBSCAP), test="F")
Single term deletions
```

Model:

```
HWFAT ~ AGE + HT + WT + ABS + TRICEPS + SUBSCAP
      Df Sum of Sq  RSS   AIC F value    Pr(F)
<none>                651.05 179.51
AGE      1      9.59 660.64 178.65  1.0463 0.309839
HT       1      1.61 652.66 177.70  0.1759 0.676225
WT       1      2.55 653.60 177.81  0.2777 0.599879
ABS      1     162.00 813.05 194.84 17.6669 7.549e-05 ***
TRICEPS  1      72.68 723.73 185.76  7.9264 0.006301 **
SUBSCAP  1       5.92 656.97 178.21  0.6458 0.424315
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that HT has the largest ϕ -value of 0.676225, so it is eliminated from the model:

```
> drop1(lm(HWFAT ~ AGE + WT + ABS + TRICEPS + SUBSCAP), test="F")
Single term deletions
```

Model:

```
HWFAT ~ AGE + WT + ABS + TRICEPS + SUBSCAP
      Df Sum of Sq    RSS    AIC F value    Pr(F)
<none>                652.66 177.70
AGE      1      9.88 662.54 176.87  1.0894 0.300098
WT       1     10.55 663.22 176.95  1.1643 0.284169
ABS      1    189.07 841.73 195.54 20.8580 1.996e-05 ***
TRICEPS  1     78.81 731.47 184.59  8.6941 0.004302 **
SUBSCAP  1      5.69 658.36 176.38  0.6281 0.430660
---
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Note that SUBSCAP has the largest ϕ -value of 0.430660, so it is eliminated from the model:

```
> drop1(lm(HWFAT ~ AGE + WT + ABS +TRICEPS), test="F")
Single term deletions
```

Model:

```
HWFAT ~ AGE + WT + ABS + TRICEPS
      Df Sum of Sq    RSS    AIC F value    Pr(F)
<none>                658.36 176.38
AGE      1     13.62 671.97 175.97  1.5097  0.2231
WT       1      6.83 665.19 175.18  0.7577  0.3869
ABS      1    220.99 879.35 196.95 24.5043 4.621e-06 ***
TRICEPS  1    201.77 860.12 195.23 22.3725 1.068e-05 ***
---
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Note that WT has the largest ϕ -value of 0.3869, so it is eliminated from the model:

```
> drop1(lm(HWFAT ~ AGE + ABS +TRICEPS), test="F")
Single term deletions
```

Model:

```
HWFAT ~ AGE + ABS + TRICEPS
      Df Sum of Sq    RSS    AIC F value    Pr(F)
<none>                665.19 175.18
AGE      1     37.59 702.78 177.47  4.1823 0.04441 *
ABS      1    282.90 948.08 200.82 31.4712 3.323e-07 ***
TRICEPS  1    198.89 864.08 193.59 22.1259 1.159e-05 ***
---
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The resulting model uses AGE, ABS, and TRICEPS to predict HWFAT.

(b) Forward selection assumes a model with an intercept only and adds the most significant (smallest ϕ -values) variables one at a time. The function `add1()` in R is used as the ϕ -values at each step are shown:

```
> add1(lm(HWFAT~1), scope=(~.+ AGE + HT + WT + ABS + TRICEPS + SUBSCAP),
+ test="F")
```

Single term additions

Model:

```
HWFAT ~ 1
      Df Sum of Sq    RSS    AIC F value    Pr(F)
<none>                6017.8  341.0
AGE      1      175.0 5842.8  340.7   2.2765    0.1355
HT       1      117.8 5900.0  341.4   1.5175    0.2218
WT       1     3237.6 2780.2  282.7  88.5045 2.219e-14 ***
ABS      1     5072.8  945.0  198.6 407.9929 < 2.2e-16 ***
TRICEPS  1     5056.3  961.5  199.9 399.6462 < 2.2e-16 ***
SUBSCAP  1     4939.0 1078.8  208.9 347.9456 < 2.2e-16 ***
---
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The variable ABS has the most significant (smallest) ϕ -value = $2.2e - 16$ with the largest F value= 407.9929, so it is added to the model:

```
> add1(lm(HWFAT~ABS), scope=(~.+AGE +HT +WT +TRICEPS +SUBSCAP), test="F")
```

Single term additions

Model:

```
HWFAT ~ ABS
      Df Sum of Sq    RSS    AIC F value    Pr(F)
<none>                944.96 198.57
AGE      1      80.88 864.08 193.59  7.0199 0.0098255 **
HT       1      61.60 883.36 195.31  5.2298 0.0250250 *
WT       1      43.73 901.22 196.87  3.6396 0.0602498 .
TRICEPS  1     242.17 702.78 177.47 25.8443 2.639e-06 ***
SUBSCAP  1     132.58 812.38 188.77 12.2400 0.0007904 ***
---
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The variable TRICEPS has the most significant (smallest) ϕ -value = $2.639e - 06$ with the largest F value= 25.8443, so it is added to the model:

```
> add1(lm(HWFAT~ABS+TRICEPS),scope=(~.+ AGE + HT + WT + SUBSCAP), test="F")
```

Single term additions

Model:

```
HWFAT ~ ABS + TRICEPS
      Df Sum of Sq    RSS    AIC F value    Pr(F)
<none>                702.78 177.47
AGE      1      37.59 665.19 175.18  4.1823 0.04441 *
HT       1      25.25 677.54 176.62  2.7574 0.10104
WT       1      30.81 671.97 175.97  3.3932 0.06947 .
SUBSCAP  1       2.24 700.54 179.22  0.2370 0.62782
---
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The variable AGE has the most significant (smallest) ϕ -value = 0.04441 with the largest F value= 4.1823, so it is added to the model:

```
> add1(lm(HWFAT~ABS+TRICEPS+AGE), scope=(~.+ HT + WT + SUBSCAP), test="F")
```

Single term additions

Model:

```
HWFAT ~ ABS + TRICEPS + AGE
      Df Sum of Sq    RSS    AIC F value    Pr(F)
<none>                665.19 175.18
HT      1      7.03 658.16 176.35  0.7796 0.3802
WT      1      6.83 658.36 176.38  0.7577 0.3869
SUBSCAP 1      1.97 663.22 176.95  0.2171 0.6427
```

None of the ϕ -values now meet the α_{crit} level of 0.20, so the model is complete with ABS, TRICEPS, and AGE being used to predict HWFAT. If a `summary` is done for the models where ABS, TRICEPS, and AGE are already in the model and HT, WT, or SUBSCAP were added individually, the ϕ -values would match the last column of the last `add1()` output:

```
> summary(lm(HWFAT~ABS+TRICEPS+AGE+HT))$coefficients
      Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 15.6108355  7.06886737  2.2083928 3.035723e-02
ABS          0.3701823  0.06550886  5.6508735 2.902965e-07
TRICEPS      0.4554293  0.09980055  4.5633949 1.990682e-05
AGE          -0.4236659  0.28898736 -1.4660361 1.469329e-01
HT           -0.1020099  0.11553071 -0.8829675 3.801523e-01
```

The ϕ -value for HT is $3.801523e-01=0.3802$ from the `add1()` output:

```
> summary(lm(HWFAT~ABS+TRICEPS+AGE+WT))$coefficients
      Estimate Std. Error  t value    Pr(>|t|)
(Intercept)  9.94025577  4.31017288  2.3062313 2.393916e-02
ABS          0.39968360  0.08074124  4.9501789 4.621329e-06
TRICEPS      0.46942072  0.09924414  4.7299591 1.068468e-05
AGE          -0.38382444  0.31238134 -1.2287048 2.231289e-01
WT           -0.01585418  0.01821376 -0.8704507 3.869075e-01
```

The ϕ -value for WT is $3.869075e-01=0.3869$ from the `add1()` output:

```
> summary(lm(HWFAT~ABS+TRICEPS+AGE+SUBSCAP))$coefficients
      Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 10.59636278  4.25550451  2.490037 1.504963e-02
ABS          0.33934952  0.07364815  4.607713 1.688668e-05
TRICEPS      0.42485168  0.13249227  3.206615 1.994247e-03
AGE          -0.53122920  0.26209533 -2.026855 4.633009e-02
SUBSCAP      0.06218487  0.13346571  0.465924 6.426572e-01
```

The ϕ -value for SUBSCAP is $6.426572e-01=0.6427$ from the `add1()` output.

Note that in both the forward and backward selection procedures, the same model results: (HWFAT ~ ABS + TRICEPS + AGE). This is not always the case.

(c) R_a^2 is used with R. The R package `leaps` is needed for the function `regsubsets()`. The arguments have predictors as a matrix first, then the response as a vector. The first six variables of `HSwrestler` are the predictors, while the response, HWFAT, is in column 7.

```

> library(leaps)
> a <- regsubsets(as.matrix(HSwrestler[,-c(7,8,9)]), HSwrestler[,7])
> summary(a)
Subset selection object
6 Variables (and intercept)
      Forced in Forced out
AGE          FALSE      FALSE
HT           FALSE      FALSE
WT           FALSE      FALSE
ABS          FALSE      FALSE
TRICEPS      FALSE      FALSE
SUBSCAP      FALSE      FALSE
1 subsets of each size up to 6
Selection Algorithm: exhaustive
      AGE HT  WT  ABS TRICEPS SUBSCAP
1  ( 1 ) " " " " " " "*" " " " "
2  ( 1 ) " " " " " " "*" "*" " "
3  ( 1 ) "*" " " " " "*" "*" " "
4  ( 1 ) "*" "*" " " "*" "*" " "
5  ( 1 ) "*" " " "*" "*" "*" "*"
6  ( 1 ) "*" "*" "*" "*" "*" "*"
> summary(a)$adjr2
[1] 0.8409068 0.8801014 0.8849817 0.8846381 0.8840129 0.8826699

```

The largest R_a^2 value is 0.8849817, which corresponds to the model with three predictors. The row beside the 3 shows "*" symbols for AGE, ABS, and TRICEPS, so these are the appropriate predictor variables.

(d) When using Mallows's C_p , the idea is to select the smallest C_p value less than or equal to p . In this case, the R package `leaps` and the output from `regsubsets()` gives the optimal value $C_4 = 2.541953$, so the three-predictor (plus an intercept) model using AGE, ABS, and TRICEPS is again selected:

```

> summary(a)$cp
[1] 29.051861 4.641808 2.541953 3.775400 5.175856 7.000000
> par(pty="s")
> plot(2:7, summary(a)$cp, ylim=c(2,7), xlab="p", ylab="Cp")
> abline(a=0, b=1)

```

(e) The function `stepAIC()` in the MASS package will compute models based on both AIC and BIC statistics. The argument `k` of this function will be set equal to 2 for the AIC statistic and $\ln(n)$ for the BIC statistic. The user needs to specify the scope of the model with the argument `scope=`. In this case, the scope of the model includes any of the six predictors AGE, HT, WT, ABS, TRICEPS, and SUBSCAP. For further details, see the `stepAIC()` help file. Initial and final output is shown from using `stepAIC()`. The starting AIC value is 179.51. The `stepAIC()` function adds or removes variables until it finds the smallest AIC value. The - before a variable indicates that the variable will be removed to produce the given AIC, while a + indicates the variable will be added to produce the given AIC.

```

> library(MASS) # For function stepAIC()
> reg.all <- lm(HWFAT ~ AGE + HT + WT + ABS + TRICEPS + SUBSCAP)
> mod.aic <- stepAIC(reg.all, direction="both",
+ scope=(~.+SUBSCAP+TRICEPS+ABS+WT+HT+AGE), k=2)

```

Start: AIC= 179.51

HWFAT ~ AGE + HT + WT + ABS + TRICEPS + SUBSCAP

	Df	Sum of Sq	RSS	AIC
- HT	1	1.61	652.66	177.70
- WT	1	2.55	653.60	177.81
- SUBSCAP	1	5.92	656.97	178.21
- AGE	1	9.59	660.64	178.65
<none>			651.05	179.51
- TRICEPS	1	72.68	723.73	185.76
- ABS	1	162.00	813.05	194.84

The final solution (with intermediate steps not printed here) is

Step: AIC= 175.18

HWFAT ~ AGE + ABS + TRICEPS

	Df	Sum of Sq	RSS	AIC
<none>			665.19	175.18
+ HT	1	7.03	658.16	176.35
+ WT	1	6.83	658.36	176.38
+ SUBSCAP	1	1.97	663.22	176.95
- AGE	1	37.59	702.78	177.47
- TRICEPS	1	198.89	864.08	193.59
- ABS	1	282.90	948.08	200.82

> mod.aic

Call:

lm(formula = HWFAT ~ AGE + ABS + TRICEPS)

Coefficients:

(Intercept)	AGE	ABS	TRICEPS
10.6161	-0.5331	0.3564	0.4656

The final model uses AGE, ABS, and TRICEPS as predictors.

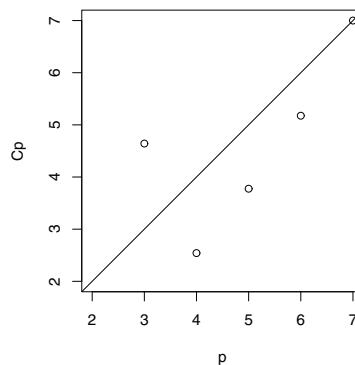


FIGURE 12.5: Plot of C_p versus p

(f) When BIC is the criterion, the model selected is $\text{HWFAT} \sim \text{ABS} + \text{TRICEPS}$. Only initial and final output are shown.

```
> mod.bic <- stepAIC(reg.all, direction="both",
+ scope=(~.+SUBSCAP+TRICEPS+ABS+WT+HT+AGE), k=log(length(HWFAT)))
Start:  AIC= 196
      HWFAT ~ AGE + HT + WT + ABS + TRICEPS + SUBSCAP
```

	Df	Sum of Sq	RSS	AIC
- HT	1	1.61	652.66	191.84
- WT	1	2.55	653.60	191.95
- SUBSCAP	1	5.92	656.97	192.35
- AGE	1	9.59	660.64	192.79
<none>			651.05	196.00
- TRICEPS	1	72.68	723.73	199.90
- ABS	1	162.00	813.05	208.98

The final model is

```
Step:  AIC=184.54
      HWFAT ~ ABS + TRICEPS
```

	Df	Sum of Sq	RSS	AIC
<none>			702.78	184.54
+ AGE	1	37.59	665.19	184.61
+ WT	1	30.81	671.97	185.40
+ HT	1	25.25	677.54	186.04
+ SUBSCAP	1	2.24	700.54	188.65
- TRICEPS	1	242.17	944.96	203.28
- ABS	1	258.75	961.54	204.64

```
> mod.bic
```

Call:

```
lm(formula = HWFAT ~ ABS + TRICEPS)
```

Coefficients:

(Intercept)	ABS	TRICEPS
2.0590	0.3371	0.5043

```
> detach(HSwrestler)
```



12.11.1.5 Summary

Variable selection is simply a means to select variables for inclusion or exclusion in a model that can be used for explanatory or predictive purposes. That is, the goal is not variable selection per se, rather, the goal is to create a model that adequately explains or predicts from the data. Stepwise selection procedures do not always guarantee a model will be selected that meets the user's need to explain or predict from the data. Criterion-based methods typically involve a wider search than do stepwise procedures, and many argue that they return models that are better than those from stepwise procedures. Regardless of the methods one uses to select a model, additional factors such as the cost to measure the variables and model diagnostics should be considered in developing a model.

12.11.2 Diagnostics

While fitting a model using the principle of least squares regression requires no distributional assumptions, using the model for inferential purposes does depend on specific assumptions. If (12.7) assumes $\varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$, that is, the errors in the model are assumed to be independent and to follow a normal distribution with a mean of zero and a constant variance, then (12.7) is called the normal error model. Regression diagnostics play a critical role in the verification of these assumptions. Regression diagnostics are also used to learn about unusual observations. The diagnostics will often dictate changes in the model selected initially. These changes emphasize the fact that model building is an iterative process.

12.11.2.1 Checking Error Assumptions

The assumption in the normal error model deals with an unobservable quantity ε . However, the residuals $\hat{\varepsilon}_i$ can be computed and analyzed. While the residuals do not have the same properties as the errors (ε), the differences between residuals and errors are slight, and examining the residuals is a reasonable approach to use in checking the assumptions about the models' errors.

First, the errors from model (12.7) are assumed to follow a normal distribution. Simple techniques such as a histogram or a density plot of the residuals can be used to study the distribution of the residuals. However, care needs to be exercised when interpreting such graphs since histograms and density plots of data that come from a normal distribution when the sample size is small will not always look normal. Furthermore, the residuals do not have a constant variance. In fact, the variance-covariance matrix for $\hat{\varepsilon}$ is

$$\text{Var}(\hat{\varepsilon}) = \sigma^2[\mathbf{I} - \mathbf{H}], \text{ where } \mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'. \quad (12.57)$$

Proof:

$$\begin{aligned} \hat{\varepsilon} &= \mathbf{Y} - \hat{\mathbf{Y}} \\ &= \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= \mathbf{Y} - \mathbf{H}\mathbf{Y} \\ &= (\mathbf{I} - \mathbf{H})\mathbf{Y} \end{aligned}$$

which implies that

$$\begin{aligned} \text{Var}(\hat{\varepsilon}) &= (\mathbf{I} - \mathbf{H}) \text{Var}(\mathbf{Y})(\mathbf{I} - \mathbf{H})' \text{ by property 3 on page 673} \\ &= (\mathbf{I} - \mathbf{H})\sigma^2\mathbf{I}(\mathbf{I} - \mathbf{H})' \\ &= \sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})' \\ \text{Var}(\hat{\varepsilon}) &= \sigma^2(\mathbf{I} - \mathbf{H}) \text{ because } (\mathbf{I} - \mathbf{H}) \text{ is symmetric and idempotent} \quad \blacksquare \end{aligned}$$

The diagonal entry of \mathbf{H} is denoted as h_{ii} , which is referred to as the leverage. Note that the trace of \mathbf{H} is p , the number of parameters ($\beta_0, \beta_1, \dots, \beta_{p-1}$) in the linear model. (See Problem 10 on page 650.)

Example 12.15 Find the \mathbf{H} matrix and display the first five h_{ii} values for the model selected with the AIC statistic from part (e) of Example 12.14 on page 598. Verify that the sum of the h_{ii} values equals p . Recall that the variables selected were AGE, ABS, and TRICEPS.

Solution: The code to produce the answer is

```
> attach(HSwrestler)
> mod.3 <- lm(HWFAT~AGE+ABS+TRICEPS)
> # Getting the values manually...
> X <- model.matrix(mod.3)
> n <- nrow(X)
> p <- ncol(X)
> H <- X%*%solve(t(X)%*%X)%*%t(X)
> hii <- diag(H)
> hii[1:5]
      1      2      3      4      5
0.05555037 0.02273942 0.03266124 0.02786396 0.20830418
> # Extracts hatvalues in R and S-PLUS
> influence(mod.3)$hat[1:5]
      1      2      3      4      5
0.05555037 0.02273942 0.03266124 0.02786396 0.20830418
> # Verifying that sum(h_ii)=p
> sum(hii)
[1] 4
> detach(HSwrestler)
```



12.11.2.1.1 Assessing Normality and Constant Variance Although formal hypothesis tests for normality, such as the Shapiro-Wilk test, can be applied to the residuals, they lack power to detect non-normal distributions. Recall that the null hypothesis in the Shapiro-Wilk test is that the distribution is normal and the alternative is that the distribution is not normal. Consider Figure 12.6 on page 610, where models 2, 3, 5, and 6 show residuals that suggest problems with either the constant variance or the normality of the errors assumption. The second residual plot on the top row (`mod2`) shows a pattern of decreasing variability. However, when a Shapiro-Wilk test is run on the residuals to test for normality, the φ -value is only mildly significant (0.08014). The increasing variance model (`mod5`) gives a conclusive rejection of normality with a φ -value of 0.006809; however, when a Shapiro-Wilk test is run on the residuals of the two far right residual plots (non-linear relationship, `mod3` and `mod6`), the test for normality on the upper right plot is only mildly significant (0.06713) while the test for normality on the bottom right residuals plot returns a highly significant φ -value (0.01138). That is, normality cannot be conclusively ruled out based on a Shapiro-Wilk test for the residuals in the top right plot (`mod3`). Consequently, it is wiser to use a combination of graphical tests as well as hypothesis tests when studying the properties of the residuals from a particular model. Using `qqnorm()` on the residuals is a good starting point for assessing normality graphically. Other graphs one might use include, but are not limited to, histograms, boxplots, and density plots. As noted earlier, care needs to be taken when interpreting such graphs.

The assumption of constant variance is typically checked by plotting the $\hat{\varepsilon}_i$ s versus the \hat{Y}_i s. Constant variance is a reasonable assumption when the residuals are scattered in a band of constant width. When the band falls around the line $y = 0$, the regression model is appropriate. Examples of constant variance are provided in the top and bottom left (`mod1` and `mod4`) residual plots of Figure 12.6 on page 610. For models 1 and 4, the φ -values from the Shapiro-Wilk test of normality are expectedly large, 0.615 and 0.6985, respectively. Models 2 and 5 have decreasing and increasing variance, respectively. One formal large sample test for constant variance is the Breusch-Pagan test, which can be

performed with the function `bptest()` from the R package `lmtest`; however, the test has no power asymptotically (Zaman, 2000).

12.11.2.1.2 Testing Autocorrelation Whenever data are obtained in a time sequence, it is possible to have correlation (called autocorrelation) among the errors. A frequently used test for detecting autocorrelation is the Durbin-Watson test. The hypotheses of the test are specified in terms of the autocorrelation coefficient ρ , $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$, and are tested with the statistic

$$DW = \frac{\sum_{i=2}^n (\hat{\epsilon}_i - \hat{\epsilon}_{i-1})^2}{\sum_{i=1}^n \hat{\epsilon}_i^2}. \quad (12.58)$$

Small values of DW lead to the conclusion that $\rho \neq 0$ because adjacent error terms $\hat{\epsilon}_i - \hat{\epsilon}_{i-1}$ tend to be similar when the data are correlated. The package `car` has a function `durbin.watson()` that can be used to test for autocorrelation. Only the far right residual plots (`mod3` and `mod6`) of Figure 12.6 on the following page have small DW values leading to φ -values of 0. The φ -values of the Durbin-Watson test for models 1, 2, 4, and 5 are 0.626, 0.24, 0.716, and 0.324, respectively.

The commands to obtain the Durbin-Watson and Shapiro-Wilk test results for the residuals in `mod1` of Figure 12.6 on the next page follow. Test results for models 2–6 stored in objects `mod2` through `mod6` can be obtained similarly.

```
> durbin.watson(mod1)
lag Autocorrelation D-W Statistic p-value
  1      0.01009468      1.923398    0.626
Alternative hypothesis: rho != 0
> shapiro.test(resid(mod1))
```

Shapiro-Wilk normality test

```
data: resid(mod1) W = 0.9894, p-value = 0.615
```

Scatterplots of residuals versus a time, sequence, or order variable can often detect non-independence of error terms. When a linear model is created and stored in an object with `S`, the function `plot()` can be applied to the linear model object and several diagnostic plots will appear on the screen. The diagnostic plots drawn for R and S-PLUS are not the same. Figure 12.7 on the following page shows the four default graphs produced with R using the function `plot()` for `mod1`. The plot in the upper left panel shows residuals plotted against fitted values. This plot can be used to detect lack of fit. If the residuals show some curvilinear trend, the current model is not appropriate; however, transforming one or more of the variables can often remedy this problem. In this graph, such a problem does not exist. The same plot can be used to assess the constant variance assumption on the errors. In this case, the variance appears constant as the fitted values vary. The second default graph is a normal quantile-quantile plot of the residuals (upper right corner of Figure 12.7). In this case, there is not a clear deviation from normality. The lower left graph plots the square root of the residuals versus the fitted values. Assuming symmetry of the errors, this graph helps assess the constant variance of the errors, which in this case seems to be a reasonable assumption. The lower right panel shows standardized residuals (as defined in (12.59)) versus leverage points. Contours for Cook's distance (as defined in (12.64)) of 0.5 and 1 facilitate an understanding of the relationship among the residuals, leverage values, and Cook's distance. R will actually produce six diagnostic graphs, but they must be specified using the argument `which=1:6`, where the `1:6` is a vector with any or all of the values 1 through 6.

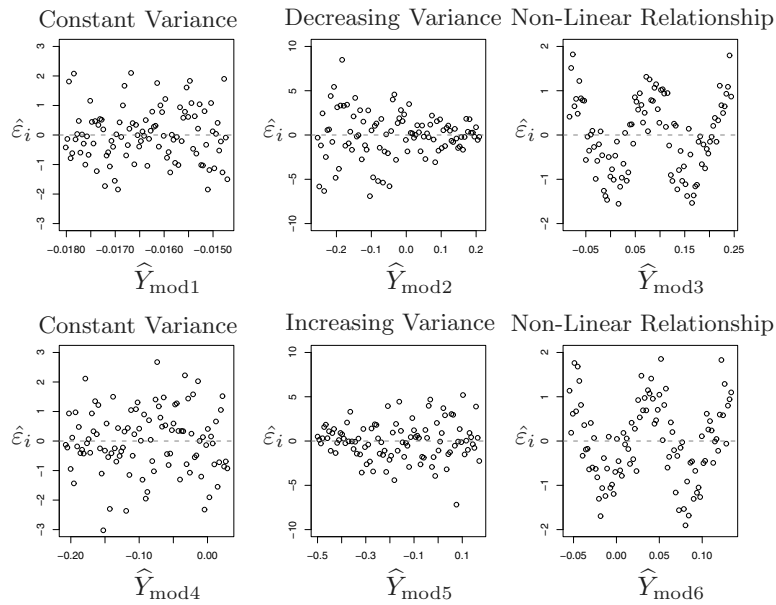


FIGURE 12.6: Residual plots for six different models with different residual patterns

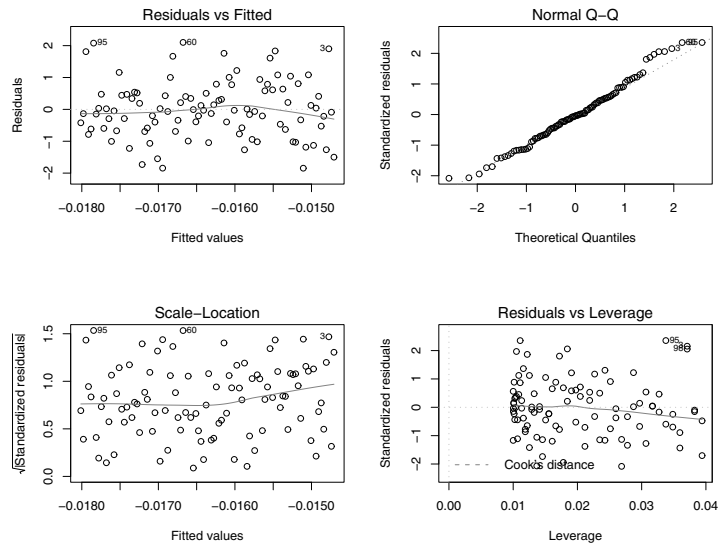


FIGURE 12.7: Diagnostic plots for mod1 in Figure 12.6

12.11.2.2 Identifying Unusual Observations

Quite often in regression models, certain observations do not seem to fit the overall pattern of the data. These cases may have a large residual and have the potential to alter dramatically the fitted regression model. An observation may be an outlier with respect to its Y values, its x values, or both, yet not all outlying observations will have a dramatic

impact on the fitted regression model. One of the ways used to measure outlying Y values is to evaluate standardized residuals. This is done because residuals may have substantially different variances. Consequently, it makes sense to consider $\hat{\varepsilon}_i$ relative to its estimated standard deviation. When the residuals are rescaled to have unit variance, the resulting residuals (r_i) are known as internally studentized residuals or **standardized residuals**, where

$$r_i = \frac{\hat{\varepsilon}_i}{\sqrt{\widehat{\text{Var}}(\hat{\varepsilon}_i)}} = \frac{\hat{\varepsilon}_i}{\hat{\sigma} \cdot \sqrt{1 - h_{ii}}}. \quad (12.59)$$

The R function `rstandard()` computes standardized residuals according to (12.59). The function `stdres()` in the MASS package (which can be used with either S-PLUS or R) also computes standardized residuals according to (12.59). Standardized residuals are sometimes preferred in residual plots since they have been standardized to have unit variance; however, in many cases, no appreciable difference will be seen between the raw residuals and the standardized residuals. Only when there is an unusually large leverage (large is generally taken to be 2 or 3 times p/n) will differences be noticeable. When standardized residuals are displayed in a quantile-quantile plot, because the residuals are standardized, the points should fall along the line $y = x$ if the normality assumption is reasonable. Figure 12.8 shows a quantile-quantile plot of the standardized residuals from the model shown in the upper left (`mod1`) of Figure 12.6 on the facing page.

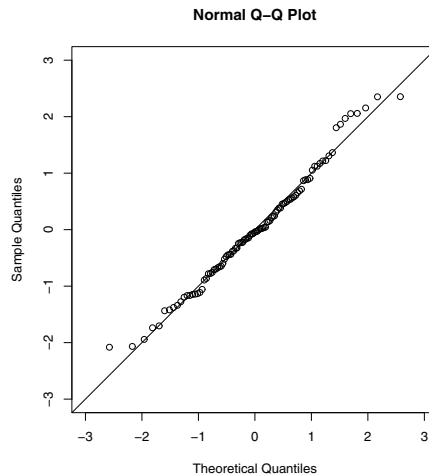


FIGURE 12.8: Quantile-quantile plot for `mod1` in Figure 12.6 on the preceding page

Another refinement to make the residuals more effective in detecting outlying observations is to use deleted residuals. Specifically, when a regression model is computed where the i^{th} case is excluded, the i^{th} prediction is denoted $\widehat{Y}_{i(i)}$, and the deleted residual ($\hat{\varepsilon}_{(i)}$) is then defined as

$$\hat{\varepsilon}_{(i)} = Y_i - \widehat{Y}_{i(i)}. \quad (12.60)$$

Fortunately, an algebraic equivalent expression for $\hat{\varepsilon}_{(i)}$ exists that does not require the computation of $\widehat{Y}_{i(i)}$ for each omitted case. Specifically, it can be shown that $\hat{\varepsilon}_{(i)} = Y_i -$

$\widehat{Y}_{i(i)} = \frac{\hat{\varepsilon}_i}{1-h_{ii}}$. The estimated variance of the $\hat{\varepsilon}_{(i)}$ is

$$\widehat{Var}[\hat{\varepsilon}_{(i)}] = \frac{\hat{\sigma}_{(i)}^2}{1-h_{ii}} = \frac{MSE_{(i)}}{1-h_{ii}} \quad (12.61)$$

Ordinarily, one prefers to study the **studentized deleted residuals** (r_i^*) rather than the ordinary deleted residuals. The i^{th} studentized deleted residual is defined as

$$r_i^* = \frac{\hat{\varepsilon}_{(i)}}{\sqrt{\widehat{Var}(\hat{\varepsilon}_{(i)})}} = \frac{\frac{\hat{\varepsilon}_i}{1-h_{ii}}}{\sqrt{\frac{\hat{\sigma}_{(i)}^2}{1-h_{ii}}}} = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)} \cdot \sqrt{1-h_{ii}}} \quad (12.62)$$

Again, there is an algebraic equivalent to (12.62) that avoids doing n regressions. The algebraic equivalent definition of r_i^* is

$$r_i^* = r_i \cdot \left(\frac{n-p-1}{n-p-r_i^2} \right)^{\frac{1}{2}} \sim t_{n-p-1} \quad (12.63)$$

When the model is correct, each studentized deleted residual follows a t -distribution with $n-p-1$ degrees of freedom. Even though it is very likely only a few “large” r_i^* s will be of interest, by identifying them as large, all cases have implicitly been tested. To control the overall significance level, a Bonferroni approach is often used where r_i^* values are declared significant if their absolute value exceeds $t_{1-\alpha/2n; n-p-1}$. However, this approach does tend to be conservative, especially for large n . The R function `rstudent()` computes the studentized deleted residuals according to (12.63). The function `studres()` in the MASS package (which can be used with R or S-PLUS) also computes the studentized deleted residuals according to (12.63).

Example 12.16 Compute and plot the residuals, standardized residuals, and studentized residuals for the model $\text{HWFAT} \sim \text{ABS} + \text{TRICEPS}$ versus the fitted values using the data frame `HSwrestler`. What HWFAT values do the residuals indicate are unusual? Can any of the studentized residuals be considered an outlier according to the Bonferroni approach if the significance level is 0.20?

Solution: The commands to calculate the solution are

```
> attach(HSwrestler)
> library(MASS)
> mod.2 <- lm(HWFAT~ABS+TRICEPS)
> par(mfrow=c(2,2))
> plot(fitted(mod.2), resid(mod.2), ylim=c(-10,10), main="")
> title(main="Residuals vs Fitted")
> abline(h=0, lty=2)
> plot(fitted(mod.2), stdres(mod.2), ylim=c(-3.5,3.5), main="")
> title(main="Standardized Residuals vs Fitted")
> abline(h=0, lty=2)
> plot(fitted(mod.2), studres(mod.2), ylim=c(-3.5,3.5), main="")
> title(main="Studentized Residuals vs Fitted")
> abline(h=0, lty=2)
> plot(mod.2, which=1, main="Default Graph 1")
> par(mfrow=c(1,1))
```

```

> sort(abs(resid(mod.2)))[76:78] # Extract three largest values
    42    22    35
6.555825 7.449100 9.495697
> sort(abs(stdres(mod.2)))[76:78] # Extract three largest values
    42    22    35
2.163508 2.458597 3.129513
> sort(abs(studres(mod.2)))[76:78] # Extract three largest values
    42    22    35
2.219409 2.546944 3.333868
> qt(1-.2/(2*78),78-3-1) # Critical value
[1] 3.121816
> detach(HSwrestler)

```

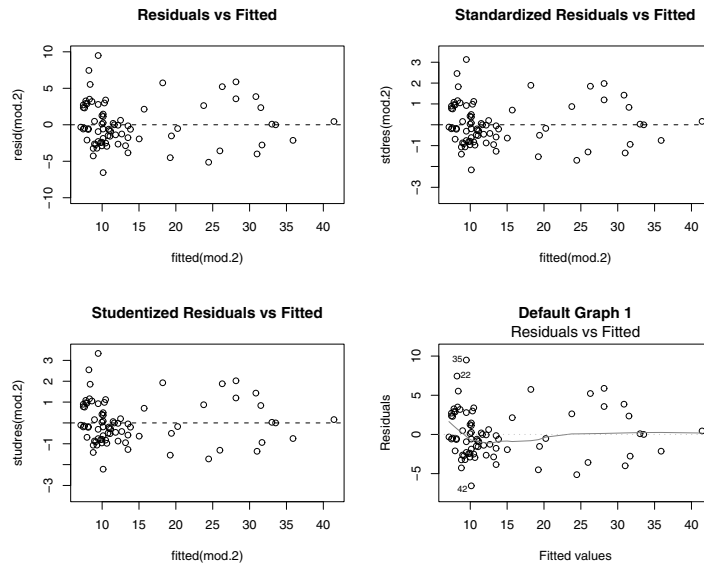


FIGURE 12.9: Residuals versus fitted values for the model $\text{HWFAT} \sim \text{ABS} + \text{TRICEPS}$

Cases 42, 22, and 35 have the largest absolute values of their plain residuals, standardized residuals, and studentized residuals. Case 35 could be considered an outlier using a significance level of $\alpha = 0.20$ since the critical value is 3.121816. ■

12.11.2.3 High Leverage Observations

While residuals were used to identify outlying Y values, the hat matrix provides an analog for the x values. The diagonal entry h_{ii} of the hat matrix \mathbf{H} provides a measure of the distance of the i^{th} case from the centroid of the x observations. That is, h_{ii} can be used to assess whether an observation is outlying from the other x s by examining its h_{ii} value. The limits on h_{ii} are $1/n \leq h_{ii} \leq 1/c$, where c is the number of rows of \mathbf{X} that have the same values as the i^{th} row. Note that the upper limit is never greater than 1. In general, a leverage value, h_{ii} , is considered large if it is more than twice as large as the mean leverage value ($2p/n$). Observations with large h_{ii} are called high leverage points, and each case should be investigated to see if the point estimates in the model under consideration change when the i^{th} case is included versus excluded from the analysis. It is important to note that not all points with high leverage will dramatically alter the estimation of parameters in the

model. When the estimated parameters are substantially different with and without the i^{th} case, the i^{th} case is said to be **influential**. That is, not all high leverage observations are influential. Clearly, which cases are influential (if any) may change when the model is changed.

Influential Observations Some influence measures examined next, all of which measure the effect of deleting the i^{th} observation, include: Cook's distance, D_i , which measures the effect on the $\hat{\beta}$ or, equivalently, on the predicted values (see (12.64)); DFFITS $_i$, which measures the effect on the predicted \hat{Y}_i s; and DFBETAS $_{k(i)}$, which measures the effect on the $\hat{\beta}_j$ s. Fortunately, all of the influence measures considered can be computed from the results of a single regression using all of the data.

Cook's Distance Cook's distance evaluates the influence of the i^{th} case on all of the n fitted values. It is a combined measure of the standardized residual (r_i) and the leverage value (h_{ii}) that produces a number used to assess the impact of removing the i^{th} observation on the all regression coefficients (β). Cook's D_i is defined as

$$\frac{(\hat{\mathbf{Y}}_{(i)} - \hat{\mathbf{Y}})'(\hat{\mathbf{Y}}_{(i)} - \hat{\mathbf{Y}})}{p\hat{\sigma}^2} = \frac{(\hat{\beta}_{(i)} - \hat{\beta})'(\mathbf{X}'\mathbf{X})(\hat{\beta}_{(i)} - \hat{\beta})}{p\hat{\sigma}^2} \quad (12.64)$$

An algebraically equivalent expression for D_i is

$$D_i = \frac{\hat{\varepsilon}_i^2}{p\hat{\sigma}^2} \left[\frac{h_{ii}}{(1 - h_{ii})^2} \right] = \frac{r_i^2}{p} \left(\frac{h_{ii}}{1 - h_{ii}} \right). \quad (12.65)$$

D_i values are generally flagged for further scrutiny when they exceed $f_{0.50;p,n-p}$; however, the exact distribution of D_i is unknown, and the use of $f_{0.50;p,n-p}$ is only a suggestion. Oftentimes, a simple graph of the D_i s will indicate values that require further scrutiny. One can always program a function according to (12.65) to compute the D_i s; however, a better approach is to use built-in functions on linear model objects. In R, `cooks.distance()` will compute the D_i s. The package `car` also has the function `cookd()` which will work in both R and S-PLUS. The function `lm.influence()` computes basic quantities used in many diagnostics, including h_{ii} values and coefficients used to compute DFBETAS. In R, `lm.influence()` returns $\hat{\beta}_{k(i)} - \hat{\beta}_k$, while in S-PLUS, $\hat{\beta}_{k(i)}$ is returned. The user should consult the documentation for further details.

DFFITS A measure related to D_i is DFFITS, which is an abbreviation for “difference in fits.” DFFITS is a standardized measure of the amount by which the predicted value \hat{Y}_i changes when the i^{th} case is deleted from the data. The definition of DFFITS is

$$\text{DFFITS}_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{h_{ii}}}, \quad (12.66)$$

while a computationally equivalent definition of DFFITS is

$$\text{DFFITS}_i = r_i^* \left(\frac{h_{ii}}{1 - h_{ii}} \right)^{\frac{1}{2}}, \quad (12.67)$$

where r_i^* is the studentized deleted residual. DFFITS values whose absolute value exceeds $2 \cdot \sqrt{p/n}$ generally require further scrutiny. To compute DFFITS with R, use `dffits(linear model)`.

It bears pointing out that there are n D_i values and n DFFITS values. The next influence measure considered is DFBETAS, which measures the influence of the i^{th} case on each regression coefficient. That is, there will be np DFBETAS values.

DFBETAS A standardized measure of the amount by which the k^{th} regression coefficient changes when the i^{th} observation is omitted from the data set is DFBETAS. A case is considered to have a large DFBETAS value if its absolute value exceeds $2/\sqrt{n}$. The DFBETAS measure is defined as

$$\text{DFBETAS}_{k(i)} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\sqrt{\hat{\sigma}_{(i)}^2 \cdot v_{k+1, k+1}}} \quad (12.68)$$

If the matrix \mathbf{V} is defined to be $(\mathbf{X}'\mathbf{X})^{-1}$, then $\sigma_{\hat{\beta}_k}^2 = \sigma^2 \cdot v_{k+1, k+1}$, where $v_{k+1, k+1}$ is the $(k+1)^{\text{st}}$ diagonal entry ($k = 0, 1, \dots, p-1$) of \mathbf{V} . To compute DFBETAS with R, use `dfbetas(linear model)`. The function `dfbetas(linear model)` in the package `car` will work for both R and S-PLUS.

Table 12.7: Summary of measures of influential observations

Influence Measure	Formula	Case i May Be Influential if:
Cook's D_i	$\frac{r_i^2}{p} \left(\frac{h_{ii}}{1 - h_{ii}} \right)$	$D_i > f_{0.5; p, n-p}$
DFFITS	$r_i^* \left(\frac{h_{ii}}{1 - h_{ii}} \right)^{\frac{1}{2}}$	$ \text{DFFITS} > 2\sqrt{\frac{p}{n}}$
$\text{DFBETAS}_{k(i)}$	$\frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\sqrt{\hat{\sigma}_{(i)}^2 \cdot v_{k+1, k+1}}}$	$ \text{DFBETAS}_{k(i)} > \frac{2}{\sqrt{n}}$

Example 12.17 \triangleright *Kinder* \triangleleft The data frame `Kinder` contains the height in inches and weight in pounds of 20 children from a kindergarten class. Use all 20 observations and construct a regression model where the results are stored in the object `mod` by regressing height on weight.

- Create a scatterplot of height versus weight to verify a possible linear relationship between the two variables.
- Compute and display the hat values for `mod` in a graph. Use the graph to identify the two largest hat values. Superimpose a horizontal line at $2p/n$. Remove the values that exceed $2p/n$ and regress height on weight, storing the results in an object named `modk`.
- Remove case 19 from the original data frame `Kinder` and regress height on weight, storing the results in `modk19`. Is the child with the largest hat value an influential observation if one considers the 19 observations without case 19 from the original data frame? Compute and consider Cook's D_i , DFFITS_i , and $\text{DFBETAS}_{k(i)}$, in reaching a conclusion. Specifically, produce a graph showing h_{ii} , the differences in $\hat{\beta}_{1(i)} - \hat{\beta}_1$, $\text{DFBETAS}_{k(i)}$, studentized residuals, DFFITS_i , and Cook's D_i along with the corresponding values that flag observations for further scrutiny assuming $\alpha = 0.10$.
- Remove case 20 from the data frame `Kinder` and regress height on weight, storing the results in `modk20`. Is the child with the largest hat value an influential observation if one considers the 19 observations without case 20 from the original data frame? Compute and consider Cook's D_i , DFFITS_i , and $\text{DFBETAS}_{k(i)}$ in reaching a conclusion. Specifically, produce a graph showing h_{ii} , the differences in $\hat{\beta}_{1(i)} - \hat{\beta}_1$, $\text{DFBETAS}_{k(i)}$,

studentized residuals, $DFFITS_i$, and Cook's D_i along with the corresponding values that flag observations for further scrutiny assuming $\alpha = 0.10$.

- (e) Create a scatterplot showing all 20 children. Use a solid circle to identify case 19 and a solid triangle to identify case 20. Superimpose the lines for models `mod` (type=1), `modk` (type=2), `mod19` (type=3), and `mod20` (type=4).

Solution: The code given is for R.

- (a) Based on Figure 12.10 created by entering

```
> attach(Kinder)
> plot(wt, ht)
```

assuming a linear relationship between height and weight appears reasonable; however, two points will bear further scrutiny.

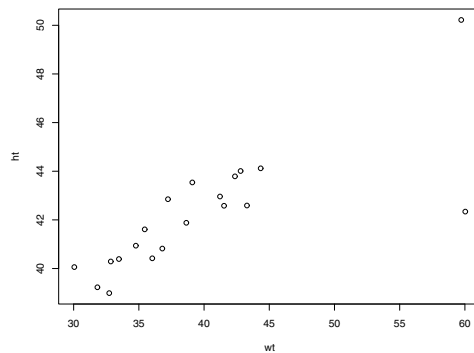


FIGURE 12.10: Scatterplot of height (`ht`) versus weight (`wt`) for the data set `Kinder`

- (b) Note that the largest h_{ii} values are for observations 19 and 20. Child 19, although taller and heavier than the other children, seems to follow the linear trend of increased height with increased weight. Child 20 appears to be right around the 50% percentile in height but has the largest weight (obese child).

```
> mod <- lm(ht~wt)
> hii <- lm.influence(mod)$hat
> hii
```

	1	2	3	4	5	6
	0.06738101	0.08955925	0.12575694	0.08161811	0.05184540	0.06981955
	7	8	9	10	11	12
	0.05268211	0.06474060	0.06038889	0.06100160	0.05773825	0.05030394
	13	14	15	16	17	18
	0.05499910	0.05573322	0.05093256	0.05688131	0.10054318	0.08821112
	19	20				
	0.37485962	0.38500423				

The following code creates the \mathbf{H} matrix and extracts the diagonal values (leverage values). It is better to use internal functions rather than matrix multiplications.

```

> X <- model.matrix(mod)
> n <- nrow(X)
> p <- ncol(X)
> H <- X%*%solve(t(X)%*%X)%*%t(X)
> hi <- diag(H)
> plot(hi, type="h", ylab="leverage")
> abline(h=2*p/n)

```

Note that observations 19 and 20 have leverage values that exceed $2p/n = 0.20$. Observations 19 and 20 are removed from consideration and `ht` is regressed on `wt` with the results stored in `modk`.

```

> modk <- lm(ht[-c(19,20)]~wt[-c(19,20)])

```

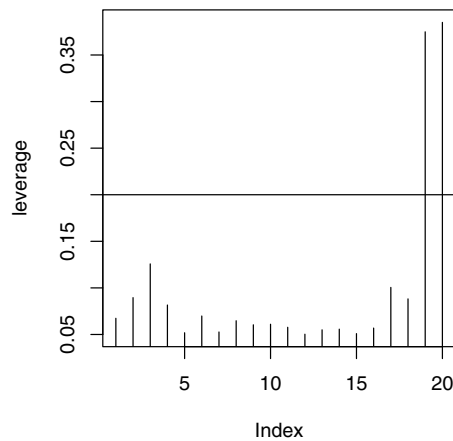


FIGURE 12.11: Graph of leverage values versus order for regressing height on weight for the data set `Kinder`

(c) The 19th observation now corresponds to the “obese” child. From the diagnostics below, the obese child is flagged in each graph for further scrutiny. Note that `lm.influence(linear model)$coefficients` returns $\hat{\beta}_{k(i)} - \hat{\beta}_k$ in R, while the same command in S-PLUS returns $\hat{\beta}_{k(i)}$. The “obese” child is an observation with high leverage that is also influential.

```

> library(MASS) # Need for function studres()
> library(car) # Need for function cookd
> modk19 <- lm(ht[-19]~wt[-19])
> n <- 19
> p <- 2
> par(mfrow=c(2,3))
> hiik19 <- lm.influence(modk19)$hat # extracting hii values
> plot(hiik19, ylab="Leverage")
> cv <- 2*p/n
> abline(h=cv, lty=2)
> plot(lm.influence(modk19)$coefficients[,2],
+ ylab="Difference in Coefficients")

```

```

> plot(dfbetas(modk19)[,2], ylab="DFBETAS")
> cv <- 2/sqrt(n) # Critical value for DFBETAS
> abline(h=c(-cv, cv), lty=2)
> plot(studres(modk19), ylab="Studentized Residuals")
> cv <- qt(1-.10/(2*n), n-p-1) # Critical value
> abline(h=c(-cv, cv), lty=2)
> DFFITS <- studres(modk19)*(hiik19/(1-hiik19))^.5 #See *
> plot(DFFITS, ylab="DFFITS")
> cv <- 2*sqrt(p/n) # Critical value for DFITS
> abline(h=c(-cv, cv), lty=2)
> cd <- cookd(modk19) # Cook's distance
> plot(cd, ylab="Cook's Distance")
> CF <- qf(.50, p, n-p) # Critical value for Cook's Distance
> abline(h=CF, lty=2)
> par(mfrow=c(1,1))

```

* DFFITS is obtained with (12.67).

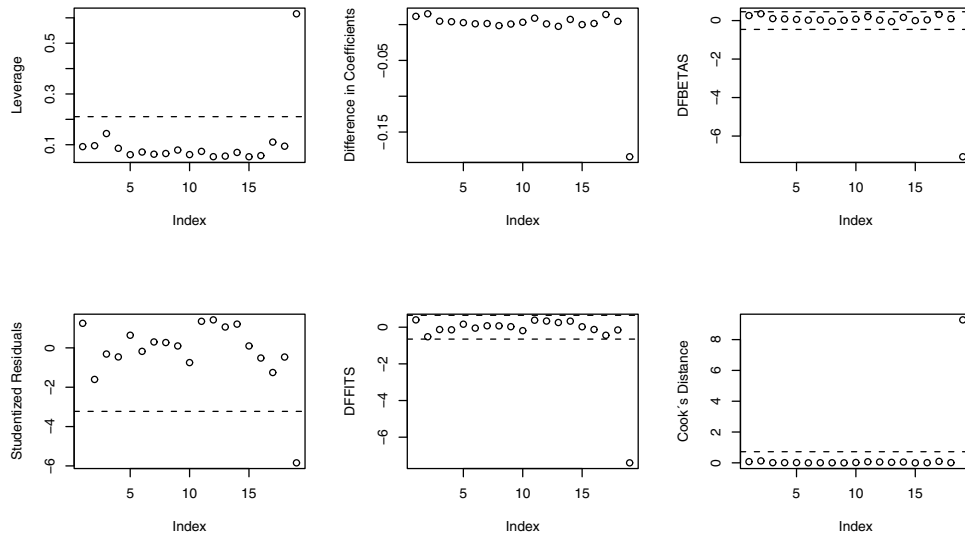


FIGURE 12.12: Diagnostic graphs for modk19 requested in part (c) of Example 12.17

(d) The 19th observation now corresponds to the “tall but normal weight” child. From the diagnostics below, this child is flagged in the leverage, DFBETAS, and DFFITS graphs for further scrutiny. Interestingly, it is not flagged with Cook’s D_i . The “tall but normal weight” child only marginally alters the regression line. Consequently, the 19th observation has high leverage but is not that influential. See the graph for part (e) for a visual explanation.

```

> modk20 <- lm(ht[-20]~wt[-20])
> n <- 19
> p <- 2
> par(mfrow=c(2,3))
> hiik20 <- lm.influence(modk20)$hat

```

```

> plot(hiik20, ylab="Leverage")
> cv <- 2*p/n
> abline(h=cv, lty=2)
> plot(lm.influence(modk20)$coefficients[,2], ylab="Difference in
+ Coefficients")
> plot(dfbetas(modk20)[,2], ylab="DFBETAS")
> cv <- 2/sqrt(n) # Critical value for DFBETAS
> abline(h=c(-cv, cv), lty=2)
> plot(studres(modk20), ylab="Studentized Residuals")
> cv <- qt(1-.10/(2*n), n-p-1) # Critical value
> abline(h=c(-cv, cv), lty=2)
> DFFITS <- studres(modk20)*(hiik20/(1-hiik20))^.5
> plot(DFFITS, ylab="DFFITS")
> cv <- 2*sqrt(p/n) # Critical value for DFITS
> abline(h=c(-cv, cv), lty=2)
> cd <- cookd(modk20) # Cook's distance
> plot(cd, ylab="Cook's Distance")
> CF <- qf(.50, p, n-p) # Critical value for Cook's Distance
> abline(h=CF, lty=2)
> par(mfrow=c(1,1))

```

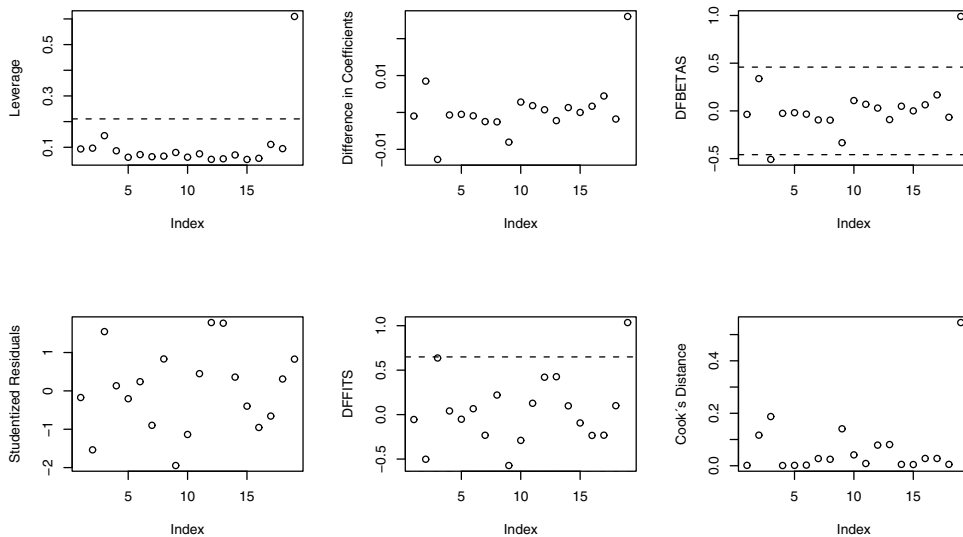


FIGURE 12.13: Diagnostic graphs for `modk20` requested in part (d) of Example 12.17

(e) In Figure 12.14 on the next page, when all 20 cases are included in the regression, cases 19 (solid circle) and 20 (solid triangle) both have large leverage values; however, if case 20 is omitted, case 19 still has a large leverage value, yet it is not very influential. Consider the differences between the lines `modk20` (dot-dash, case 20 omitted) and `modk` (dash, where cases 19 and 20 are omitted). There is very little difference between the lines `modk20` and `modk`. On the other hand, if case 19 (solid circle) is omitted, the resulting regression `modk19` (dotted) is substantially different from `modk`. In other words, case 20 has high leverage and is influential when case 19 is omitted.

```

> plot(wt[-c(19,20)], ht[-c(19,20)], cex=2, xlim=c(28,62), ylim=c(36,52),
+ xlab="Weight in Pounds", ylab="Height in Inches")
> abline(mod, lty=1, lwd=2)
> abline(modk, lty=2, lwd=2)
> abline(modk19, col="red", lty=3, lwd=2)
> abline(modk20, col="blue", lty=4, lwd=2)
> abline(mod, lty=4, lwd=2)
> points(wt[19], ht[19], pch=16, cex=2, col="red")
> points(wt[20], ht[20], pch=17, cex=2, col="blue")
> detach(Kinder)

```

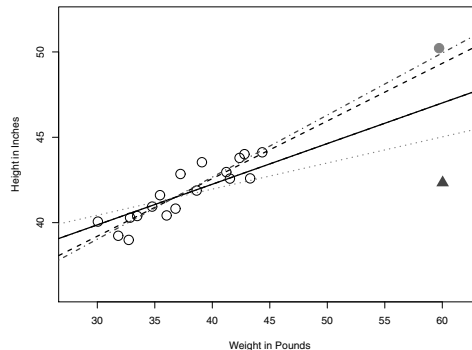


FIGURE 12.14: Scatterplot of height versus weight for data from `Kinder` with four superimposed regression lines: The solid line is for model `mod` (all observations); the dot-dash line is for model `modk20` (case 20 omitted); the dashed line is for model `modk` (case 19 and 20 omitted); and the dotted line is for model `modk19` (case 19 omitted). ■

12.11.3 Transformations

When residual analysis reveals serious problems, or when the relationships between the response and predictors are clearly non-linear, regression may still yield a reasonable model with either a transformation of the response variable, the predictors, or both response and predictors. When a scatterplot between the response and a predictor shows a non-linear relationship where the residuals are reasonably normal in distribution, appropriate transformations on the predictor may linearize the relationship between the variables without drastically altering the distribution of the residuals. After the transformation of the predictor(s), the residuals produced with the transformed variable(s) in the new model will need to be reanalyzed to assure normality assumptions are still satisfied.

Example 12.18 ▷ *Transformation of Predictors* ◁ The data frame `SimDataXT` contains simulated data for the response, Y , and predictors, x_1 , x_2 , and x_3 . Apply appropriate transformations to x_1 , x_2 , and x_3 to linearize the relationships between the response and predictors one at a time.

Solution: The answers, computed with R, are

Transform x_1 : The top left graph in Figure 12.15 shows a non-linear relationship between Y and x_1 . The second graph shows the residuals from regressing Y on x_1 , both the first and second graphs suggest a simple transformation on x_1 . The pattern suggests a square root transformation. The resulting scatterplot and residual analysis for regressing Y on $x_1^{0.5} = \sqrt{x_1}$ are illustrated in the bottom row of graphs. The curvilinear relationship evident in both the scatterplot and the residual plot using the untransformed x_1 disappear once a square root transformation is applied to x_1 .

```
> attach(SimDataXT)
> par(mfrow=c(2,3))
> plot(x1, Y)
> lines(x1, x1^.5)           # function Y = x1^.5
> plot(lm(Y~x1), which=c(1,2)) # Residual and Q-Q normal plots
> plot(x1^.5, Y)
> mod1 <- lm(Y~I(x1^.5))    # Works in R for S-PLUS see *
> abline(mod1)
> plot(mod1, which=c(1,2))  # Q-Q plot in S-PLUS is 4 not 2
> par(mfrow=c(1,1))
```

The identity function, $I()$, is used to inhibit the interpretation of \sim as a formula operator. Operators such as $+$, $-$, $*$, and \wedge have different meanings in formulas. In cases where the user wants to use arithmetical operators in a formula, they should be protected with the identity function.

* There are negative values in x_1 , and taking their square root produces NA values. R, by default, removes missing observations in its $lm()$ function with `na.action=na.omit`; however, S-PLUS does not. To remove the NA observations while using the function $lm()$ in S-PLUS, use the argument `na.action=na.exclude` inside the $lm()$ function.

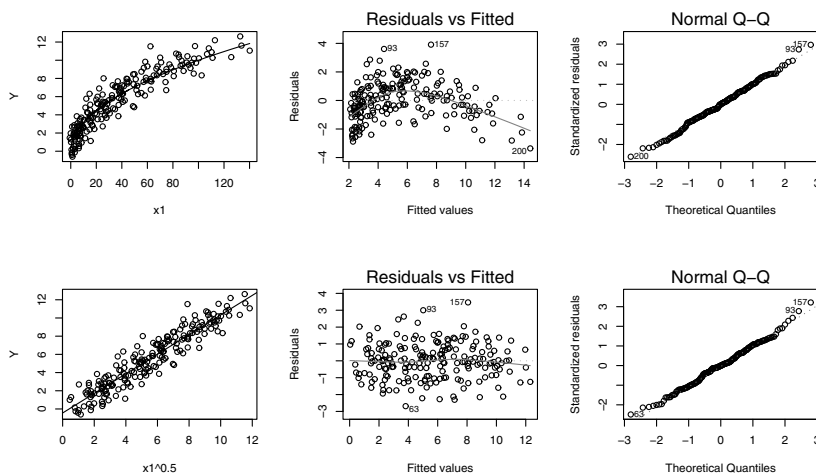


FIGURE 12.15: Scatterplot, residuals versus fitted values, and quantile-quantile plot of standardized residuals for Y versus x_1 and Y versus $x_1^{0.5}$ models

Transform x_2 : The concave up relationship depicted in the first two graphs of Figure 12.16 suggests a quadratic transformation on x_2 . The resulting scatterplot and residual graphs for the transformed predictor are depicted in the bottom row of graphs.

```
> par(mfrow=c(2,3))
> plot(x2, Y)
> lines(x2, x2^2) # function Y = x2^2
> plot(lm(Y~x2),which=c(1,2)) # Residuals and Q-Q normal plots
> plot(x2^2, Y)
> mod2 <- lm(Y~I(x2^2))
> abline(mod2)
> plot(mod2, which=c(1,2))
> par(mfrow=c(1,1))
```

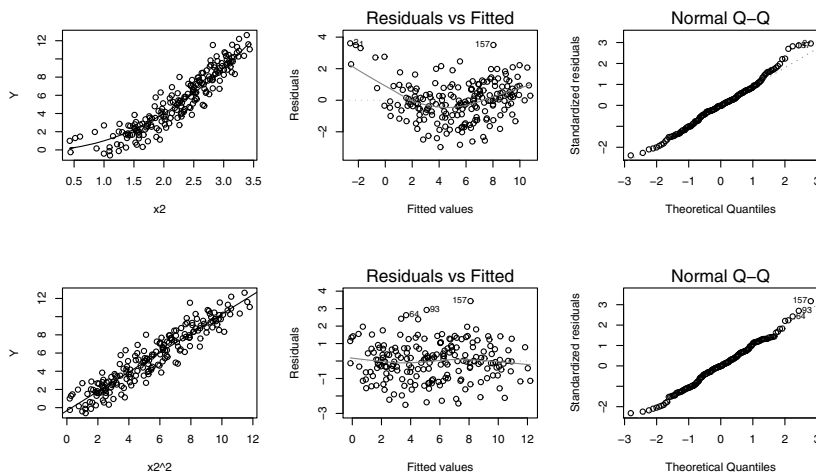


FIGURE 12.16: Scatterplot, residuals versus fitted values, and quantile-quantile plot of standardized residuals for Y versus x_2 and Y versus x_2^2 models

Transform x_3 : The first two graphs of Figure 12.17 on the facing page suggest a reciprocal transformation on x_3 . As before, the graphs in the second row of Figure 12.17 are for the transformed predictor (x_3).

```
> par(mfrow=c(2,3))
> plot(x3, Y)
> lines(x3, x3^(-1)) # function Y = 1/x3
> plot(lm(Y~x3),which=c(1,2)) # Residuals and Q-Q normal plots
> plot(x3^(-1), Y)
> mod3 <- lm(Y~I(x3^(-1)))
> abline(mod3)
> plot(mod3, which=c(1,2))
> par(mfrow=c(1,1))
> detach(SimDataXT)
```

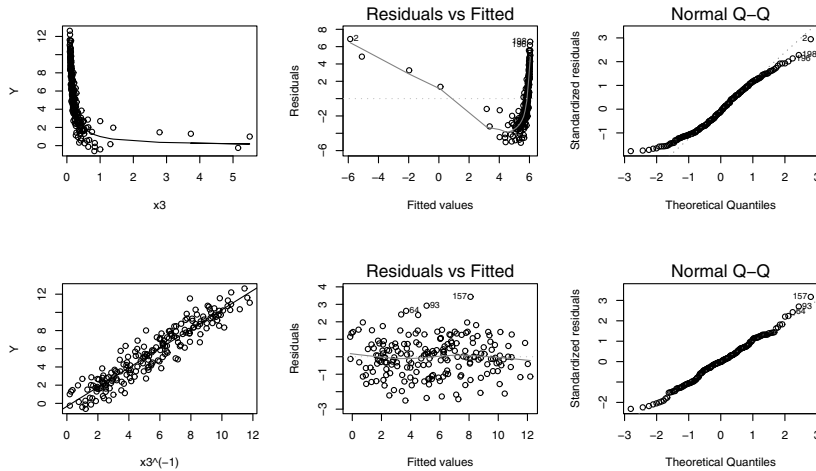


FIGURE 12.17: Scatterplot, residuals versus fitted values, and quantile-quantile plot of standardized residuals for Y versus x_3 and Y versus x_3^{-1} models ■

12.11.3.1 Collinearity

Collinearity in regression occurs when some of the predictors are a linear combination of other predictors. When $\mathbf{X}'\mathbf{X}$ is singular, there is said to be exact collinearity and there is no unique estimate of β . When $\mathbf{X}'\mathbf{X}$ is near singular, the problem is often called **multicollinearity**. Multicollinearity causes problems with the estimation of β and its subsequent interpretation. Severe multicollinearity can cause the sign of the coefficients to be opposite what is expected and typically inflates the standard errors of the estimates to the point where variables appear no longer to be significant. Two techniques to detect collinearity include computation of the condition number and computation of the variance inflation factor.

The **condition number** κ is defined as the square root of the largest eigenvalue of $\mathbf{X}'\mathbf{X}$ divided by the smallest eigenvalue of $\mathbf{X}'\mathbf{X}$. κ values between 30 and 100 indicate that there are moderate to strong dependencies among the predictors. κ values greater than 100 indicate serious multicollinearity problems. The S function `kappa()` can be used to estimate the condition number of a matrix.

A related method of detecting multicollinearity is to regress x_j on all of the other predictors. When the coefficient of determination (R_j^2) from regressing x_j on all of the other predictors is near one, there is multicollinearity among the predictors. The **variance inflation factor** is defined as

$$\text{VIF}_j = \frac{1}{1 - R_j^2}. \quad (12.69)$$

When there are dependencies among the predictors, R_j^2 will be near one and VIF_j will be large. VIF_j values greater than 10 suggest serious collinearity. The VIF_j for a predictor x_j can be interpreted as the factor ($\sqrt{\text{VIF}_j}$) by which the standard error of $\hat{\beta}_j$ is increased due to the presence of multicollinearity.

Example 12.19 ▷ *Multicollinearity* ◁ In Example 12.18 on page 620, using the data frame `SimDataXT`, Y was regressed on the transformed variables x_1 , x_2 , and x_3 one at a time.

- (a) Regress Y on $x_1^{0.5}$, x_2^2 , and x_3^{-1} and store the results in the object `modC`. Are there any linear dependencies among the predictors?

- (b) Regress Y on $x_1^{0.5}$ and x_2^2 and store the results in the object `modB`. Compute the condition number for `modB` and the VIF for $x_1^{0.5}$ and x_2^2 . Verify that the standard error for $\hat{\beta}_1$ from a model where Y is regressed solely on $x_1^{0.5}$ (`mod1`) and the standard error for $\hat{\beta}_1$ from `modB` increases by approximately $\sqrt{\text{VIF}_1}$.

Solution: The answers are as follows:

- (a) From the output it is seen that $\mathbf{X}'\mathbf{X}$ is singular. In particular, x_2 is a function of x_3 ($x_2 \equiv \frac{1}{\sqrt{x_3}}$). The output shown is from R. S-PLUS will not compute coefficients for a singular model.

```
> attach(SimDataXT)
> modC <- lm(Y~I(x1^.5)+I(x2^2)+I(x3^(-1)))
> summary(modC)
```

Call:

```
lm(formula = Y ~ I(x1^0.5) + I(x2^2) + I(x3^(-1)))
```

Residuals:

	Min	1Q	Median	3Q	Max
	-2.56394	-0.77548	-0.01170	0.75323	3.43862

Coefficients: (1 not defined because of singularities)

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.4219	0.1724	-2.447	0.0153 *
I(x1^0.5)	0.4500	0.5412	0.831	0.4068
I(x2^2)	0.6244	0.5384	1.160	0.2475
I(x3^(-1))	NA	NA	NA	NA

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.082 on 195 degrees of freedom

Multiple R-Squared: 0.8861, Adjusted R-squared: 0.8849

F-statistic: 758.2 on 2 and 195 DF, p-value: < 2.2e-16

```
> cbind(x2, x3^(-.5))[1:5,]
```

	x2	
[1,]	0.5177716	0.5177716
[2,]	0.4262969	0.4262969
[3,]	0.4405244	0.4405244
[4,]	0.5988884	0.5988884
[5,]	0.8465002	0.8465002

- (b) The condition number for `modB` is 87.53672, suggesting strong dependencies exist between $x_1^{0.5}$ and x_2^2 . The variance inflation factor for $x_1^{0.5}$ and x_2^2 is 382.7241. Finally, the standard error for $\hat{\beta}_1$ based on `mod1` is 0.0277 while the standard error for $\hat{\beta}_1$ based on `modB` is 0.5412. The ratio of 0.5412 to 0.0277 is 19.54, which is approximately equal to the square root of the VIF for `modB` (19.56). In this problem, the introduction of x_2^2 to a model that already contained $x_1^{0.5}$ increased the standard error for $\hat{\beta}_1$ by 19.56. From the summary of `modB`, neither of the estimated coefficients for β_1 or β_2 are significant, yet from Example 12.18 on page 620, the coefficients for both $x_1^{0.5}$ and x_2^2 , when taken alone, are significant.

Note that the output shown is from R. To remove the NA observations (produced from taking the square root of a negative value) while using the function `lm()` in S-PLUS, use the option `na.action=na.exclude` inside the function `lm()`.

```
> modB <- lm(Y~I(x1^.5)+I(x2^2))
> X <- model.matrix(modB)
> eigen(t(X)%*%X, only.values=TRUE)$values # extracting eigenvalues
[1] 15394.239140  39.566738  2.008989
> lambda.max <- max(eigen(t(X)%*%X, only.values=TRUE)$values)
> lambda.min <- min(eigen(t(X)%*%X, only.values=TRUE)$values)
> condition.number <- (lambda.max/lambda.min)^.5
> condition.number
[1] 87.53672
```

Verify the results with the function `kappa()`. The argument `exact=TRUE` used with the function `kappa()` only works with R.

```
> kappa(X, exact=TRUE)
[1] 87.53672
```

Compute the VIF with (12.69):

```
> 1/(1-summary(lm(I(x1^.5)~I(x2^2)))$r.square)
[1] 382.7241
```

Compute the variance inflation factors with the function `vif()` from the `car` package:

```
> library(car) # For function vif()
> vif(modB)
I(x1^0.5)  I(x2^2)
 382.7241  382.7241
```

Verify that the standard error for $\hat{\beta}_1$ from a model where Y is regressed solely on $x_1^{0.5}$ to `modB` increases by approximately $\sqrt{\text{VIF}_1}$:

```
> mod1 <- lm(Y~I(x1^.5))
> summary(mod1)$coefficients
      Estimate Std. Error  t value    Pr(>|t|)
(Intercept) -0.4410092  0.17178289 -2.567247 1.099639e-02
I(x1^0.5)    1.0768822  0.02769023 38.890330 4.239490e-94
> se.x1.mod1 <- summary(mod1)$coefficients[2,2]
> se.x1.modB <- summary(modB)$coefficients[2,2]
> ratio <- se.x1.modB/se.x1.mod1
> ratio
[1] 19.54613
> vif(modB)^.5
I(x1^0.5)  I(x2^2)
 19.56334  19.56334
> detach(SimDataXT)
```



12.11.3.2 Transformations for Non-Normality and Unequal Error Variances

With the normal distribution, the mean and variance are independent of one another. This is not the case with many other distributions. One such example is the Poisson distribution, where the mean is equal to the variance. Quite often, non-normality and unequal error variances appear together. This “double” problem can often be remedied by transforming the response variable \mathbf{Y} . The “double” problem can be identified by an increasing or decreasing band in a curvilinear residual plot. Transformations on the response variable will frequently both linearize a curvilinear relationship and fix the problem of unequal error variances. Other times, transformations on both the response and predictors will be required to meet the assumptions of the normal linear model. One technique that searches computationally for an appropriate transformation of the response variable that directly addresses normality is the Box-Cox method. The Box-Cox method estimates the parameter λ for the transformation $Y' = Y^\lambda$, where

$$Y' = \begin{cases} \frac{Y^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0, \\ \ln Y & \text{for } \lambda = 0, \end{cases} \quad (12.70)$$

by the method of maximum likelihood. The function `boxcox()` of the `MASS` package produces a plot of the log-likelihood against the transformation parameter λ for a particular model. By default, the range of λ is from -2 to 2 . However, once the value of λ that maximizes the log-likelihood is known, the range of the plot in `boxcox()` can be tightened to highlight the area where the function is maximized with the argument `lambda=`. For more details, see the `boxcox()` help file. The `boxcox()` function is generally used just to get an idea for an appropriate transformation. The value of λ that maximizes the log-likelihood function may turn out to be 0.53 ; but if there is a possible explanation for taking the square root of the response, the transformation applied should be $\lambda = 0.5$ and not the value that maximizes the log-likelihood function.

Example 12.20 ▷ *Box-Cox Transformation* ◁ Use the data frame `SimDataST` and the `boxcox()` function to find the transformation on Y that maximizes the log-likelihood of the model created by regressing Y_1 on x_1 . Once the value of λ that maximizes the log-likelihood is known, reduce the range of the plot produced with `boxcox()` to focus on the area around the value of λ that maximizes the log-likelihood.

Solution: Using the default range $-2 < \lambda < 2$, the `boxcox()` function shows that the transformation $\lambda = 0$, that is, $\ln Y$, comes close to maximizing the log-likelihood and is included in the 95% confidence band for λ , as seen in Figure 12.18 on the next page. Consequently, the range of λ is reduced and plotted over the region -0.3 to 0.3 using the argument `lambda=seq(-.3, .3, .01)`:

```
> attach(SimDataST)
> library(MASS)
> par(mfrow=c(1,2)) # 1 row by 2 columns
> modx1 <- lm(Y1~x1)
> boxcox(modx1)
> boxcox(modx1, lambda=seq(-.3,.3,.01))
> par(mfrow=c(1,1))
> detach(SimDataST)
```

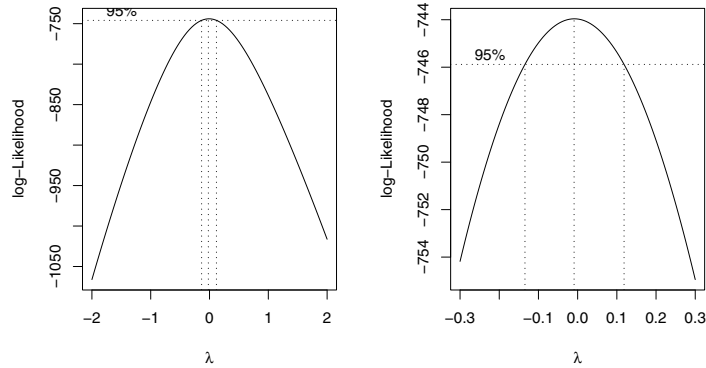


FIGURE 12.18: Box-Cox graph of λ for Example 12.20 on the facing page ■

Example 12.21 ▷ *Transforming Y with \ln* ◁ Use the data frame `SimDataST` to create and display six graphs using a 2 by 3 layout. Start by producing a scatterplot of Y_1 versus x_1 . Plot the residuals versus fits for the model created by regressing Y_1 on x_1 (call this model `modx1`). Based on the first two graphs, does a logarithmic transformation for the response variable make sense? Use and plot the results from the `boxcox()` function applied to `modx1`. In the second row of graphs, create a scatterplot of $\ln Y_1$ versus x_1 , a plot of the residuals versus the fits for the model $\log(Y_1) \sim x_1$, and a quantile-quantile normal plot of the residuals from the model $\log(Y_1) \sim x_1$. Based on the second row of graphs, do the assumptions for the normal error model seem to be satisfied for the model $\log(Y_1) \sim x_1$? Note that the default understanding of `log` in both R and S-PLUS is $\log_e = \ln = \log$.

Solution: Based on the first two graphs of the first row of Figure 12.19 on the next page, and the subsequent plot of λ , the transformation $\lambda = 0$, that is, $\ln(Y)$, is justified. Once the response is transformed, the problems of non-normality and unequal variance of the errors apparently disappear.

```
> attach(SimDataST)
> library(MASS)
> par(mfrow=c(2,3))
> plot(x1, Y1)
> modx1 <- lm(Y1~x1)
> plot(modx1, which=1)
> boxcox(modx1, lambda=seq(-.3,.3,.01))
> plot(x1, log(Y1))
> plot(lm(log(Y1)~x1), which=c(1,2)) #Q-Q plot in S-PLUS is 4 not 2
> par(mfrow=c(1,1))
> detach(SimDataST)
```

Example 12.22 ▷ *Transforming Y with a Reciprocal* ◁ Use the data frame `SimDataST` to create and display six graphs using a 2 by 3 layout. Start by producing a scatterplot of Y_2 versus x_2 . Plot the residuals versus fits for the model created by regressing Y_2 on x_2 (call this model `modx2`). Based on the first two graphs, does a reciprocal

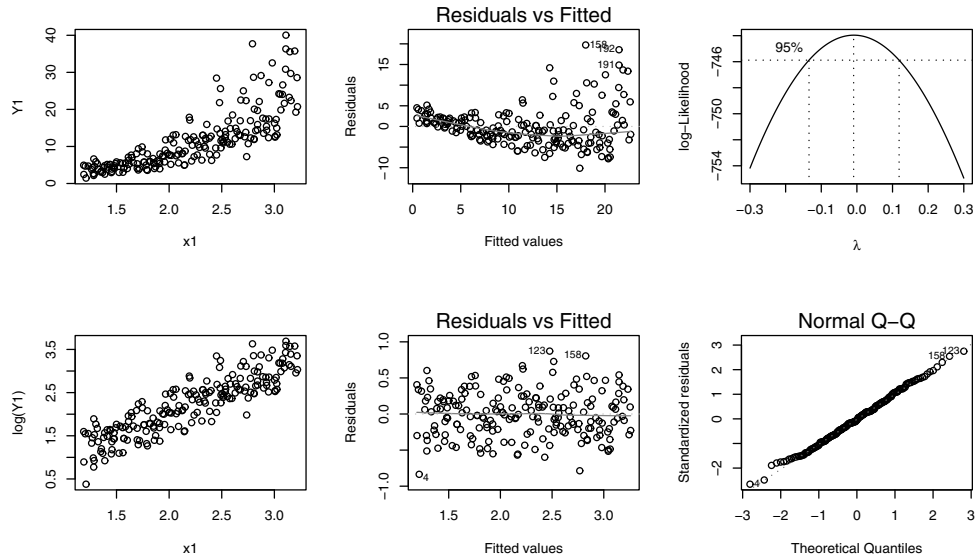


FIGURE 12.19: Scatterplot and residual versus fitted plot of Y_1 versus x_1 ; Box-Cox plot of λ ; scatterplot, residual versus fitted plot, and quantile-quantile plot of $\ln(Y_1)$ versus x_1

transformation for the response variable make sense? Use and plot the results from the `boxcox()` function applied to `modx2`. In the second row of graphs, create a scatterplot of Y_2^{-1} versus x_2 , a plot of the residuals versus the fits for the model $I(Y_2^{-1}) \sim x_2$, and a quantile-quantile normal plot of the residuals from the model $I(Y_2^{-1}) \sim x_2$. Based on the second row of graphs, do the assumptions for the normal error model seem to be satisfied for the model $I(Y_2^{-1}) \sim x_2$?

Solution: Based on the first two graphs of the first row of Figure 12.20 on the facing page, and the subsequent plot of λ , the transformation $\lambda = -1$, that is, Y_2^{-1} , is justified. Once the response is transformed, the non-normality problem as well as the unequal variance of the errors problem appear to vanish.

```
> attach(SimDataST)
> par(mfrow=c(2,3))
> plot(x2, Y2)
> modx2 <- lm(Y2~x2)
> plot(modx2, which=1)
> boxcox(modx2, lambda=seq(-1.4,-.6,.01))
> plot(x2, Y2^(-1))
> plot(lm(I(Y2^(-1))~x2), which=c(1,2)) #Q-Q plot in S-PLUS is 4 not 2
> par(mfrow=c(1,1))
```



As mentioned earlier, at times it will be necessary to transform the response and the predictor. Consider the top left graph in Figure 12.21 on page 630 along with the bottom left graph produced with `boxcox()` suggesting a $\ln Y$ transformation. The middle column of graphs depicts a scatterplot of $\log_e Y$ versus x_3 with a superimposed line from the least

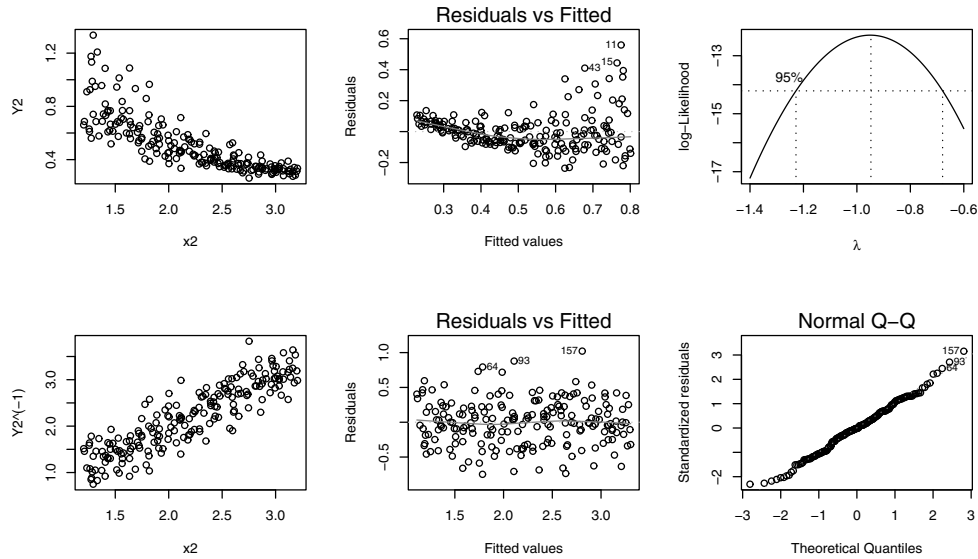


FIGURE 12.20: Scatterplot and residual versus fitted plot of Y_2 versus x_2 ; Box-Cox plot of λ ; scatterplot, residual versus fitted plot, and quantile-quantile plot of Y_2^{-1} versus x_2

squares fit of the model $\log(Y_1) \sim x_3$ as well as the residuals versus fits plot for the same model. Close scrutiny reveals a slight curvilinear pattern in both of the graphs in the middle column. This suggests some type of transformation for x_3 . The slight curvature is eliminated in both the scatterplot and the residual plot by applying a square root transformation to x_3 . The R code used to create Figure 12.21 is

```
> library(MASS)
> par(mfrow=c(2,3))
> plot(x3, Y1)
> plot(x3, log(Y1))
> mod <- lm(log(Y1)~x3)
> abline(mod)
> plot(x3^.5, log(Y1))
> mod2 <- lm(log(Y1)~I(x3^.5))
> abline(mod2)
> boxcox(lm(Y1~x3), lambda=seq(-.3,.3,.01))
> plot(mod, which=1)
> plot(mod2, which=1)
> par(mfrow=c(1,1))
> detach(SimDataST)
```

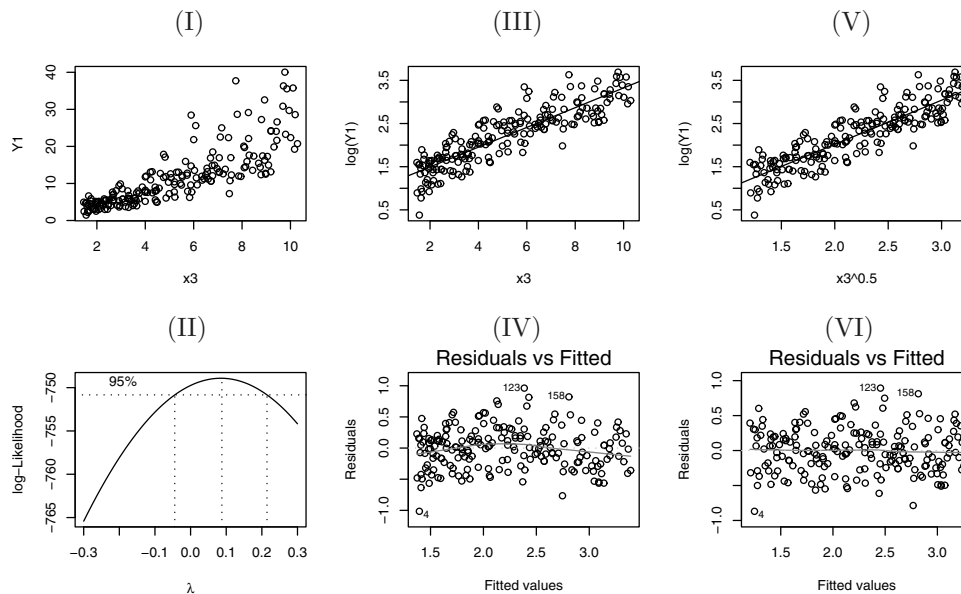



FIGURE 12.21: Process of model building with transformations: (I) original scatterplot, (II) `boxcox()` transformation suggestion, (III) scatterplot with Y transformed, (IV) residual plot shows curvature, (V) x -variable transformed, (VI) residuals appear normalized

12.12 Interpreting a Logarithmically Transformed Model

Variables are often transformed to fix constant variance or normality assumptions; however, transformations can complicate the interpretation of the model. Unlike many other transformations, models that use logarithmic transformations have approximate explanations without back transforming the variables.

When x has been transformed with a natural log transformation, the change in the $\ln(x)$ is roughly equal to the change in x provided the changes in x are small. Consider Figure 12.22 on the facing page, which graphically illustrates how changing the x values 3 and 6 by 10% corresponds to an approximate increase in $\ln(x)$ of about 10%.

An example from economics that has multiplicative error terms is the demand function ($Q = \alpha P^\beta \varepsilon$), where Q = quantity demanded, P = price, α and β are unknown parameters, and ε is the error term. This function is often transformed by taking the natural logarithm of both sides. That is,

$$\ln(Q) = \ln(\alpha) + \beta \ln(P) + \ln(\varepsilon) \quad (12.71)$$

which is in the form of a simple linear model ($Y = \beta_0 + \beta_1 x + \varepsilon$). Note that the errors in a simple linear model are additive.

The parameter β in (12.71) can be interpreted as the percent change in Q over the percent change in P , which is the definition of **price elasticity**. In other words, $|\beta|$ = price elasticity. When dealing with a simple linear model of the form

$$\ln(Y) = \ln(\beta_0) + \beta_1 \ln(x) + \varepsilon, \quad (12.72)$$

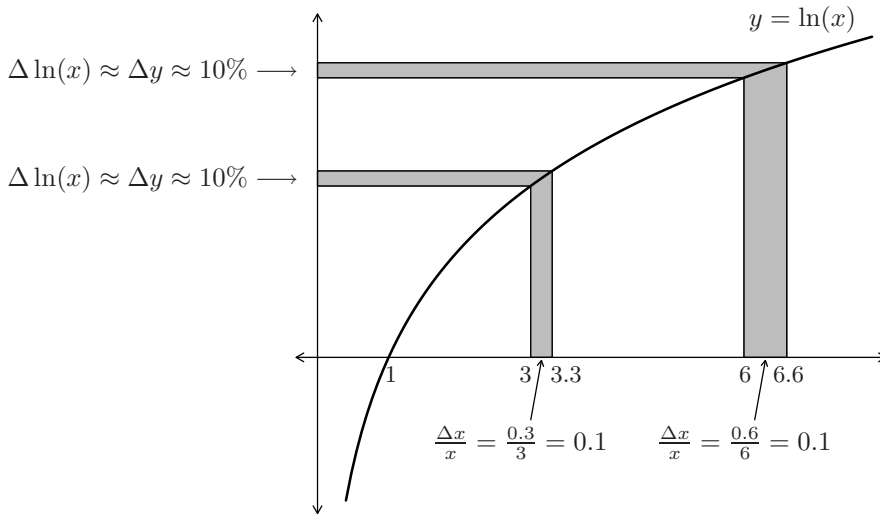


FIGURE 12.22: Small change in x gives a similar small change in $\ln(x)$

β_1 can be interpreted as

$$\beta_1 = \frac{\% \Delta y}{\% \Delta x}. \quad (12.73)$$

In Example 2.24 on page 61, the data frame **Animals** from the **MASS** package was used to find the least squares line for regressing $\log(\text{brain})$ on $\log(\text{body})$ once the three dinosaurs were removed from the data. (Note that $\log(x)$ in **S** is the natural logarithm function, $\ln(x)$.) The resulting least squares estimates of β_0 and β_1 after the dinosaurs are removed are

```
> library(MASS)
> attach(Animals)
> SA <- Animals[order(body),]
> NoDINO <- SA[-c(28:26),]
> detach(Animals)
> attach(NoDINO)
> Y <- log(brain)
> x <- log(body)
> simple.model <- lm(Y~x)
> (simple.model)$coef
(Intercept)      x
  2.1504121  0.7522607
> detach(NoDINO)
```

If the body weight of an animal increases by 1%, the approximate increase in brain weight is $(0.01 \times 0.75 = 0.0075 = 0.75\%)$ since $\hat{\beta}_1 = 0.75$. The predicted brain weight of the Jaguar whose weight is listed as 100 kg. with the fitted model is $2.15 + 0.75 \times \ln(100) = 5.61$. However, this must be back transformed to get the units of original brain measurement (grams). The brain weight predicted by the model is $\exp(5.61) = 274.431$ g. If said

Jaguar were to increase its weight by 10%, the expected increase in brain weight would be approximately 7.5% for a new weight of $1.075 \times 274.431 = 295.014$ g. The actual brain weight change predicted by the model for a body weight of 110 kg is 294.83 g and the change in brain weight as predicted from the model is 7.43% (see Table 12.8). Note that for this model, $\hat{\beta}_1 = 0.0752 \approx \Delta Y = 0.0743$. In fact, when both the response and the predictors have been transformed with a natural logarithm, one can use the percentage interpretation of β_1 as in (12.73) and be very close to the actual change given by the model for small changes in the x -variables. The parameters of growth models of the form $P(t) = ce^{\beta t}$ are

Table 12.8: Actual change in Jaguar brain weight

	x	$\ln(x)$	$\ln(Y)$	Y
	100.0	4.605	5.615	274.43
	110.0	4.700	5.686	294.83
Δ	0.1			0.0743

often estimated with ordinary least squares regression after taking the natural logarithms of both sides since $\ln P(t) = \ln(c) + \beta t$ is the form of a simple linear model. When the slope, β , is estimated for such a model, it provides an estimate of the approximate growth rate in units of t . More generally, for models of the form $\ln Y = \beta_0 + \beta_1 x$, for each unit of increase in x , Y increases roughly by $\beta_1 \times 100\%$.

12.13 Qualitative Predictors

Up to this point, only quantitative (continuous) predictor variables have been used in regression models. Quantitative variables take on values on a well-defined scale. Examples include height, weight, income, and age, to name a few; however, many predictor variables are qualitative. For example, gender (male/female) or race (Caucasian, Hispanic, Asian, etc.) are qualitative variables that appear in many regression models. Regression using quantitative variables can be generalized to qualitative variables with the use of dummy variables. A **dummy variable** is any variable in a regression model that takes on a finite number of values so that different categories of a nominal variable can be identified. Provided the regression model has an intercept, one must define $k - 1$ dummy variables to define a qualitative variable with k categories. There are many ways to define the $k - 1$ dummy variables. R and S-PLUS (Version 8) use treatment contrasts by default to define qualitative variables (factors). To see the values R or S-PLUS use to define a qualitative variable with four levels, enter

```
> contr.treatment(4)
  2 3 4
1 0 0 0
2 1 0 0
3 0 1 0
4 0 0 1
```

The rows of this matrix (4×3) are the levels of the qualitative predictor and the columns are the dummy variables. R assigns levels to a qualitative variable in alphabetical order by default.

Example 12.23 ▷ *Ease Levels Dummy Variables* ◁ Consider the variable *Ease* from the *EPIDURAL* data frame. Define appropriate dummy variables to specify the three levels of this variable.

Solution: The three levels of *Ease* (Difficult, Easy, and Impossible) require two dummy variables to be able to identify all three levels of *Ease*:

```
> attach(EPIDURAL)
> contrasts(Ease)
      Easy Impossible
Difficult  0         0
Easy      1         0
Impossible 0         1
```

Note that the first level in alphabetical order is *Difficult*. To change the first level of *Ease* to *Easy*, enter

```
> levels(Ease) <- c("Easy", "Difficult", "Impossible")
> levels(Ease)
[1] "Easy"      "Difficult"  "Impossible"
> contrasts(Ease)
      Difficult Impossible
Easy         0         0
Difficult    1         0
Impossible   0         1
> detach(EPIDURAL)
```

The simplest situation where dummy variables might be used in a regression model is when the qualitative predictor has only two levels. The regression model for a single quantitative predictor (x_1) and a dummy variable (D_1) is written

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 D_1 + \beta_3 x_1 D_1 + \varepsilon \quad (12.74)$$

where

$$D_1 = \begin{cases} 0 & \text{for the first level} \\ 1 & \text{for the second level} \end{cases}$$

The model in (12.74) when D_1 has two levels will yield one of four possible scenarios, as shown in Figure 12.23 on the next page. This type of model requires the user to answer three **basic questions**:

- (1) Are the lines the same?
- (2) Are the slopes the same?
- (3) Are the intercepts the same?

To address basic question (1), the null hypothesis $H_0 : \beta_2 = \beta_3 = 0$ must be tested. One way to perform the test is to use the general linear test statistic based on the full model found in (12.74) and the reduced model $Y = \beta_0 + \beta_1 x_1 + \varepsilon$. If the null hypothesis is not rejected, the interpretation is that there is one line present (the intercept and the

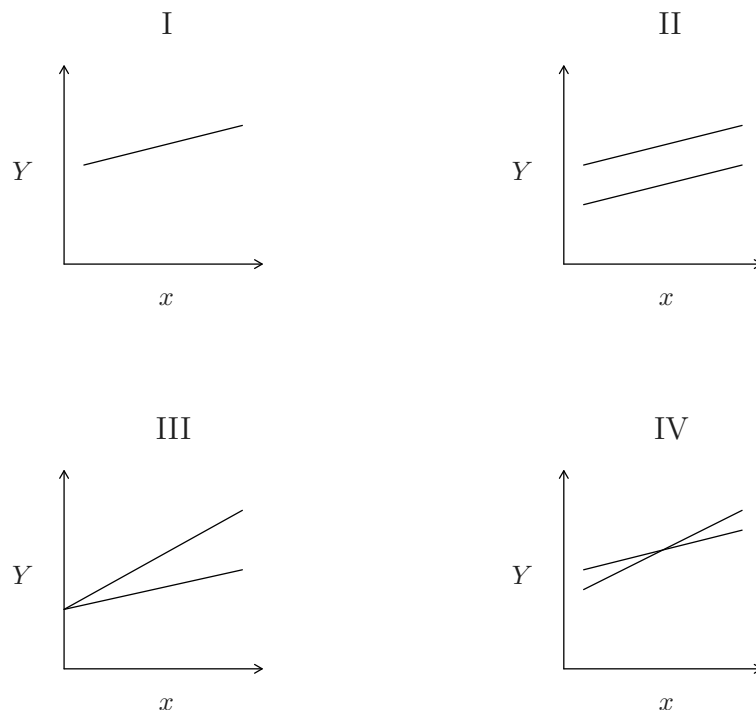


FIGURE 12.23: Four possible results for a single dummy variable with two levels. Graph I has the intercept and the slope the same for both levels of the dummy variable. Graph II has the two lines with the same slope, but different intercepts. Graph III shows the two fitted lines with the same intercept but different slopes. Graph IV shows the two lines with different intercepts and different slopes.

slope are the same for both levels of the dummy variable). This is the case for graph I of Figure 12.23. If the null hypothesis is rejected, either the slopes, the intercepts, or possibly both the slope and the intercept are different for the different levels of the dummy variable, as seen in graphs II, III, and IV of Figure 12.23, respectively.

To answer basic question (2), the null hypothesis $H_0 : \beta_3 = 0$ must be tested. If the null hypothesis is not rejected, the two lines have the same slope, but different intercepts, as show in graph II of Figure 12.23. The two parallel lines that result when $\beta_3 = 0$ are

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon \text{ for } (D_1 = 0) \quad \text{and} \quad Y = (\beta_0 + \beta_2) + \beta_1 x_1 + \varepsilon \text{ for } (D_1 = 1)$$

When $H_0 : \beta_3 = 0$ is rejected, one concludes that the two fitted lines are not parallel as in graphs III and IV of Figure 12.23.

To answer basic question (3), the null hypothesis $H_0 : \beta_2 = 0$ for model (12.74) must be tested. The reduced model for this test is $Y = \beta_0 + \beta_1 x_1 + \beta_3 x_1 D_1 + \varepsilon$. If the null hypothesis is not rejected, the two fitted lines have the same intercept but different slopes:

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon \text{ for } (D_1 = 0) \quad \text{and} \quad Y = \beta_0 + (\beta_1 + \beta_3) x_1 + \varepsilon \text{ for } (D_1 = 1)$$

Graph III of Figure 12.23 represents this situation. If the null hypothesis is rejected, one concludes that the two lines have different intercepts, as in graphs II and IV of Figure 12.23.

Example 12.24 ▷ *Elevators* ◁ Suppose a realtor wants to model the appraised price of an apartment as a function of the predictors living area (in m²) and the presence or absence of elevators. Consider the data frame `vit2005`, which contains data about apartments in Vitoria, Spain, including `totalprice`, `area`, and `elevator`, which are the appraised apartment value in Euros, living space in square meters, and the absence or presence of at least one elevator in the building, respectively.

- (a) The realtor first wants to know if there is any relationship between appraised price (Y) and living area (x_1).
- (b) Next, the realtor wants to know how adding a dummy variable for whether or not an elevator is present changes the relationship:
 - (i) Are the lines the same?
 - (ii) Are the slopes the same?
 - (iii) Are the intercepts the same?

Solution: (a) A linear regression model of the form

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon \quad (12.75)$$

is fit yielding

$$\hat{Y}_i = 40822.4 + 2704.8x_{i1}$$

and a scatterplot of `totalprice` versus `area` with the fitted regression line superimposed over the scatterplot is show in Figure 12.24 on the following page.

```
> attach(vit2005)
> Elevator <- as.factor(elevator)
> contrasts(Elevator)
  1
0 0
1 1
> modSimpl <- lm(totalprice~area)
> summary(modSimpl)
```

Call:

```
lm(formula = totalprice ~ area)
```

Residuals:

Min	1Q	Median	3Q	Max
-156126	-21564	-2155	19493	120674

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	40822.4	12170.1	3.354	0.00094 ***
area	2704.8	133.6	20.243	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 40810 on 216 degrees of freedom

Multiple R-Squared: 0.6548, Adjusted R-squared: 0.6532

F-statistic: 409.8 on 1 and 216 DF, p-value: < 2.2e-16

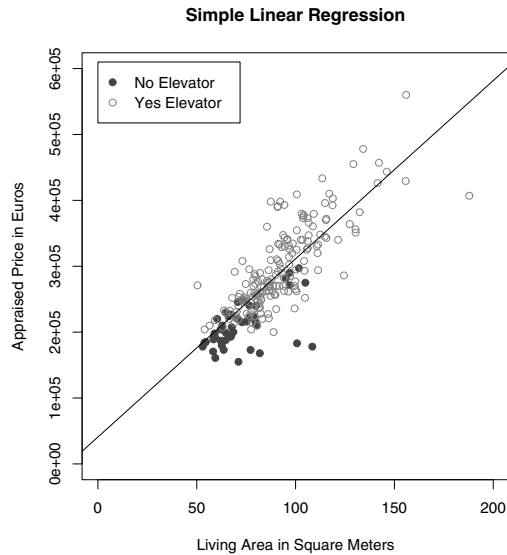


FIGURE 12.24: Scatterplot of `totalprice` versus `area` with the fitted regression line superimposed

Based on Figure 12.24, there appears to be a linear relationship between appraised price and living area. Further, this relationship is statistically significant, as the p -value for testing $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ is less than 2.2×10^{-16} .

(b) The regression model including the dummy variable for `Elevator` is written

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 D_1 + \beta_3 x_1 D_1 + \varepsilon \quad (12.76)$$

where

$$D_1 = \begin{cases} 0 & \text{when a building has no elevators} \\ 1 & \text{when a building has at least one elevator} \end{cases}$$

(i) To determine if the lines are the same (which means that the linear relationship between appraised price and living area is the same for apartments with and without elevators), the hypotheses are

$$H_0 : \beta_2 = \beta_3 = 0 \text{ versus } H_1 : \text{at least one } \beta_i \neq 0 \text{ for } i = 2, 3.$$

```
> modTotal <- lm(totalprice~area+Elevator+area:Elevator)
> modSimpl <- lm(totalprice~area)
> anova(modSimpl, modTotal)
Analysis of Variance Table
```

```
Model 1: totalprice ~ area
Model 2: totalprice ~ area + Elevator + area:Elevator
  Res.Df    RSS Df Sum of Sq    F Pr(>F)
1     216 3.5970e+11
2     214 3.0267e+11  2  5.7040e+10 20.165 9.478e-09 ***
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

In this problem, one may conclude that at least one of β_2 and β_3 is not zero since the ϕ -value = 9.478×10^{-9} . In other words, the lines have either different intercepts, different slopes, or different intercepts and slopes.

(ii) To see if the lines have the same slopes (which means that the presence of an elevator adds constant value over all possible living areas), the hypotheses are

$$H_0 : \beta_3 = 0 \text{ versus } H_1 : \beta_3 \neq 0.$$

```
> anova(modTotal)
Analysis of Variance Table
```

```
Response: totalprice
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
area	1	6.8239e+11	6.8239e+11	482.4846	< 2.2e-16 ***
Elevator	1	4.5308e+10	4.5308e+10	32.0352	4.83e-08 ***
area:Elevator	1	1.1732e+10	1.1732e+10	8.2949	0.00438 **
Residuals	214	3.0267e+11	1.4143e+09		

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the ϕ -value = 0.00438, it may be concluded that $\beta_3 \neq 0$, which implies that the lines are not parallel.

(iii) To test for equal intercepts (which means that appraised price with and without elevators starts at the same value), the hypotheses to be evaluated are

$$H_0 : \beta_2 = 0 \text{ versus } H_1 : \beta_2 \neq 0.$$

```
> modTotal <- lm(totalprice~area+Elevator+area:Elevator)
> modInter <- lm(totalprice~area+area:Elevator)
> anova(modInter, modTotal)
Analysis of Variance Table
```

```
Model 1: totalprice ~ area + area:Elevator
```

```
Model 2: totalprice ~ area + Elevator + area:Elevator
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	215	3.0624e+11				
2	214	3.0267e+11	1	3.5765e+09	2.5288	0.1133

Since the ϕ -value for testing the null hypothesis is 0.1133, one fails to reject H_0 and should conclude that the two lines have the same intercept but different slopes.

```
> summary(modInter)$coef
              Estimate Std. Error  t value    Pr(>|t|)
(Intercept)  71352.0844 12309.1794  5.796656 2.389680e-08
area         1897.9368   180.5908 10.509596 4.082287e-21
area:Elevator1 553.9856    90.4240  6.126534 4.227047e-09
> detach(vit2005)
```

The fitted model is $\hat{Y}_i = 7135 + 1898x_{i1} + 554x_{i1}D_{i1}$, and the fitted regression lines for the two values of D_1 are shown in Figure 12.25 on the following page. The fitted model using the same intercept with different slopes has an R_a^2 of 0.7034, a modest improvement over the model without the variable `Elevator`, which had an R_a^2 value of 0.6532.

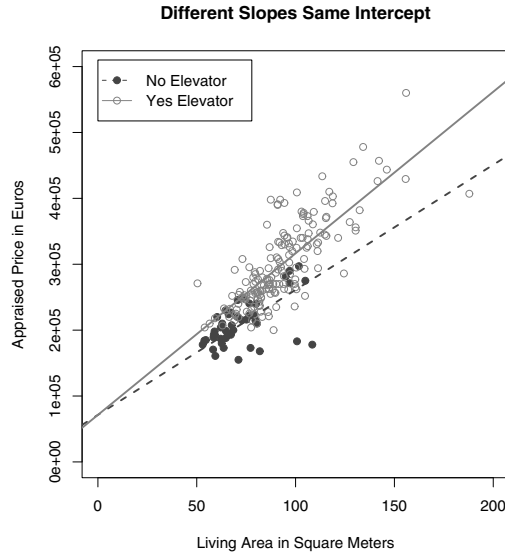


FIGURE 12.25: Fitted regression lines for Elevators example

As the numbers of levels in the qualitative variables increases, the number of dummy variables required to represent all of the possible combinations of variables (both dummy and numerical) increases rapidly, and the comparison of regression equations becomes virtually intractable. Further exploration on this topic could be carried out with a book dedicated to regression.

12.14 Estimation of the Mean Response for New Values \mathbf{X}_h

Not only is it desirable to create confidence intervals on the parameters of the regression models, but it is also common to estimate the mean response ($E(Y_h)$) for a particular set of \mathbf{X} values. The particular values where an estimate is desired will be denoted $\mathbf{X}_h = [1, x_{h,1}, x_{h,2}, \dots, x_{h,p-1}]$. Since $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, it follows that $\hat{Y}_h = \mathbf{X}_h\hat{\boldsymbol{\beta}}$. For the normal error model ($\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$),

$$\hat{Y}_h \sim N(Y_h = X_h\beta, \sigma^2\mathbf{X}_h(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_h). \quad (12.77)$$

Recall (12.32) states that $\mathbf{s}_{\hat{\boldsymbol{\beta}}}^2 = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1} = MSE(\mathbf{X}'\mathbf{X})^{-1}$, while (12.77) gives $\sigma_{\hat{Y}_h}^2 = \sigma^2\mathbf{X}_h(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_h$, from which it follows that

$$\mathbf{s}_{\hat{Y}_h}^2 = MSE \cdot \mathbf{X}_h(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_h = \mathbf{X}_h\mathbf{s}_{\hat{\boldsymbol{\beta}}}^2\mathbf{X}'_h. \quad (12.78)$$

Consequently, for a vector of given values (\mathbf{X}_h), a $(1 - \alpha) \cdot 100\%$ confidence interval for the mean response $E(Y_h)$ is

$$CI_{1-\alpha}[E(Y_h)] = \left[\hat{Y}_h - t_{1-\alpha/2; n-p} \cdot \mathbf{s}_{\hat{Y}_h}, \hat{Y}_h + t_{1-\alpha/2; n-p} \cdot \mathbf{s}_{\hat{Y}_h} \right] \quad (12.79)$$

The S function `predict()` applied to a linear model object will compute \hat{Y}_h and $s_{\hat{Y}_h}$ for a given \mathbf{X}_h . S output has \hat{Y}_h labeled `fit` and $s_{\hat{Y}_h}$ labeled `se.fit`. The function `predict()` can be used for a wide range of applications where the statistician would like to predict values of new data. One of `predict()`'s arguments is `newdata=`, where what follows the `=` should be a data frame whose columns have identical names to those of the variables that were used in constructing the original model.

12.15 Prediction and Sampling Distribution of New Observations

$Y_{h(\text{new})}$

In Section 12.14, a confidence interval was found for the mean response, $E(Y_h)$. In contrast, it is not unusual to require a confidence interval on a single, new observation instead. For example, suppose a linear model that describes course grade as a function of time studied is calculated. As the user of this model, you might be interested in your predicted grade given the amount of time you study rather than the average grade that is received by all people who study the amount of time you do. Although the point estimates for the average grade given time studied ($E(Y_h)$) and your grade given time studied ($Y_{h(\text{new})}$) are identical, the confidence intervals for these two quantities are not the same because $s_{\hat{Y}_{h(\text{new})}}^2$ accounts for an additional source of variability not present in $s_{\hat{Y}_h}^2$. Specifically, $s_{\hat{Y}_{h(\text{new})}}^2$ estimates the variance of the distribution of Y at $\mathbf{X} = \mathbf{X}_h$, which has a value of σ^2 with MSE as well as the variance of the sampling distribution of \hat{Y}_h with $s_{\hat{Y}_h}^2$.

For the normal error model,

$$\hat{Y}_{h(\text{new})} \sim N\left(Y_h = X_h\beta, \sigma^2(1 + \mathbf{X}_h(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_h)\right). \quad (12.80)$$

It follows that $s_{\hat{Y}_{h(\text{new})}}^2 = MSE(1 + \mathbf{X}_h(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_h)$ and the $(1 - \alpha) \cdot 100\%$ prediction interval for new observation $Y_{h(\text{new})}$ is written as

$$PI_{1-\alpha}[Y_{h(\text{new})}] = \left[\hat{Y}_h - t_{1-\alpha/2; n-p} \cdot s_{\hat{Y}_{h(\text{new})}}, \hat{Y}_h + t_{1-\alpha/2; n-p} \cdot s_{\hat{Y}_{h(\text{new})}}\right]. \quad (12.81)$$

To compute prediction intervals, the S function `predict()` may be applied to a linear model using the argument `interval="pred"` in R and `pi.fit=T` in S-PLUS.

Example 12.25 Use the `Grades` data set and model `gpa` as a function of `sat` assuming that the requirements for model (12.4) are satisfied.

- Compute the expected GPA (`gpa`) for an SAT score (`sat`) of 1300.
- Construct a 90% confidence interval for the mean GPA for students scoring 1300 on the SAT.
- Find the prediction limits on GPA for a future student who scores 1300 on the SAT.

Solution: The answers are as follows:

- The expected GPA for an SAT score of 1300 is $\hat{Y}_h = \mathbf{X}_h \cdot \hat{\beta} = 2.830488$, where $\mathbf{X}_h = (1, 1300)$ and $\hat{\beta} = [-1.192, 0.003]'$

```

> attach(Grades)
> Y <- gpa
> x <- sat
> simple.model <- lm(Y~x)
> betahat <- simple.model$coef
> betahat
      (Intercept)           x
-1.192063812    0.003094270
> Xh <- matrix(c(1,1300), nrow=1)
> Yhath <- Xh%*%betahat
> Yhath
      [,1]
[1,] 2.830488

```

A method that requires less typing is

```

> predict(simple.model, newdata=data.frame(x=1300))
[1] 2.830488

```

(b) A 90% confidence interval for the mean `gpa` for students scoring 1300 on the SAT using (12.79) is $CI_{0.90}(E(Y_h)) = [2.759760, 2.901216]$.

```

> MSE <- anova(simple.model)[2,3]
> MSE
[1] 0.1595551
> XTXI <- summary(simple.model)$cov.unscaled
> XTXI
      (Intercept)           x
(Intercept)  0.3101379642 -2.689270e-04
x            -0.0002689270  2.370131e-07
> var.cov.b <- MSE*XTXI
> var.cov.b
      (Intercept)           x
(Intercept)  4.948408e-02 -4.290866e-05
x            -4.290866e-05  3.781665e-08
> s2yhath <- Xh%*%var.cov.b%*%t(Xh)
> s2yhath
      [,1]
[1,] 0.001831706
> syhath <- sqrt(s2yhath)
> syhath
      [,1]
[1,] 0.04279843
> crit.t <- qt(.95,198)
> crit.t
[1] 1.652586
> CI.EYh <- c(Yhath - crit.t*syhath, Yhath + crit.t*syhath)
> CI.EYh
[1] 2.759760 2.901216

```

Or, in R only,

```
> predict(lm(Y~x), data.frame(x=1300), interval="conf", level=.90)
           fit      lwr      upr
[1,] 2.830488 2.759760 2.901216
```

Or, for S-PLUS only,

```
> predict(lm(Y~x), data.frame(x=1300), ci.fit=TRUE, conf.level=.90)$ci.fit
      lower      upper
1 2.75976 2.901216
attr(, "conf.level"):
[1] 0.9
```

(c) The prediction limits on GPA for a future student who scores 1300 on the SAT are $PI_{0.90} = [2.166595, 3.494380]$ using (12.81).

```
> s2yhathnew <- MSE + s2yhath
> syhathnew <- sqrt(s2yhathnew)
> syhathnew
      [,1]
[1,] 0.4017297
> PI <- c(Yhath - crit.t*syhathnew, Yhath + crit.t*syhathnew)
> PI
[1] 2.166595 3.494380
```

Or, in R only,

```
> predict(lm(Y~x), data.frame(x=1300), interval="pred", level=.90)
           fit      lwr      upr
[1,] 2.830488 2.166595 3.494380
> detach(Grades)
```

Only in S-PLUS:

```
> predict(lm(Y~x), data.frame(x=1300), pi.fit=TRUE, conf.level=.90)$pi.fit
      lower      upper
1 2.166595 3.494380
attr(, "conf.level"):
[1] 0.9
> detach(Grades)
```



12.16 Simultaneous Confidence Intervals

Now that a determination has been made of a correct $(1 - \alpha) \cdot 100\%$ confidence interval for a single β_k , confidence intervals for multiple β_k s are desired such that the significance level of all the intervals together will be only a specified α . For example, if $\alpha = 5\%$ and two independent confidence intervals were created for a β_0 and a β_1 , the probability that both would contain their parameters would be only $(0.95)^2 = 0.9025$, giving a family $\alpha = 0.0975$. The goal is to create intervals such that the family α is a given value. This goal is more difficult than the example because the same data are used to construct all the confidence intervals, so they are not independent and the α calculation is not straightforward. A **family confidence coefficient** is the proportion of confidence intervals that contain all the β_k parameters specified for the entire family of $g \leq p$ parameters for a given sample.

One approach to calculating these simultaneous confidence intervals is named the **Bonferroni** method. In this method, the joint interval estimates for $\beta_k, k = 0, \dots, g$ parameters are

$$CI_{1-\alpha}(\beta_k) = \left[\hat{\beta}_k - t_{1-\frac{\alpha}{2g}; n-p} \cdot s_{\hat{\beta}_k}, \hat{\beta}_k + t_{1-\frac{\alpha}{2g}; n-p} \cdot s_{\hat{\beta}_k} \right] \quad (12.82)$$

A second approach is to construct a simultaneous confidence region for the β_k coefficients. Any set of parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_g)$ that satisfy the inequality

$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\mathbf{X}'\mathbf{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{q \cdot MSE} \leq f_{1-\alpha; q, n-p} \quad (12.83)$$

fall inside a $(1 - \alpha) \cdot 100\%$ ellipsoidal confidence region for $\boldsymbol{\beta}$ where MSE is that of the full model. Note that q is the rank of \mathbf{K} for the hypothesis $H_0 : \mathbf{K}\boldsymbol{\beta} = \mathbf{m}$ discussed in Section 12.10. When the simultaneous confidence limits are for $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_k)$, excluding β_0 , q will be equal to $p - 1$, the number of predictors in the full model. This is a rather computationally intensive method. The simultaneous Scheffé confidence limits for the individual β_k s based on (12.83) are

$$CI_{1-\alpha}(\beta_k) = \left[\hat{\beta}_k - \sqrt{q \cdot f_{1-\alpha; q, n-p}} \cdot s_{\hat{\beta}_k}, \hat{\beta}_k + \sqrt{q \cdot f_{1-\alpha; q, n-p}} \cdot s_{\hat{\beta}_k} \right]. \quad (12.84)$$

For simple linear regression, the function `confidence.ellipse()` in the `car` package will construct and display a simultaneous confidence region for β_0 and β_1 ($q = 2$). For models with $p > 2$, `confidence.ellipse()` will draw a simultaneous confidence region for any two β_k s, $k = 1, \dots, p - 1$, specified by the user. Note that in the $p > 2$ case, q will equal $p - 1$. The R function `confint()` will compute individual confidence intervals for one or more parameters in a fitted model. The Bonferroni intervals from (12.82) will be wider than those from (12.84) whenever $t_{1-\frac{\alpha}{2g}; n-p} > \sqrt{q \cdot f_{1-\alpha; q, n-p}}$.

12.16.1 Simultaneous Confidence Intervals for Several Mean Responses — Confidence Band

To construct several confidence intervals for the mean response, $E(Y_h)$, corresponding to different \mathbf{X}_h vectors such that the family confidence coefficient is $1 - \alpha$, use (12.85), where

$$s_{\hat{Y}_h} = \sqrt{\mathbf{X}_h s_{\hat{\beta}}^2 \mathbf{X}_h'}.$$

$$CI_{1-\alpha}[E(Y_h)] = \left[\hat{Y}_h - \sqrt{p \cdot f_{1-\alpha; p, n-p}} \cdot s_{\hat{Y}_h}, \hat{Y}_h + \sqrt{p \cdot f_{1-\alpha; p, n-p}} \cdot s_{\hat{Y}_h} \right] \quad (12.85)$$

or

$$CI_{1-\alpha}[E(Y_h)] = \left[\hat{Y}_h - t_{1-\frac{\alpha}{2g}; n-p} \cdot s_{\hat{Y}_h}, \hat{Y}_h + t_{1-\frac{\alpha}{2g}; n-p} \cdot s_{\hat{Y}_h} \right] \quad (12.86)$$

whichever produces narrower intervals.

A **confidence band** is a region of confidence around the entire regression line constructed by plotting the upper and lower values of (12.85) over the range of \mathbf{X}_h and subsequently connecting all of the upper values with a curve and all of the lower values with a curve. See Figure 12.26 on the next page for an example.

12.16.2 Predictions of g New Observations

To create simultaneous prediction intervals for g new observations, corresponding to g \mathbf{X}_h vectors with a family confidence coefficient of $1 - \alpha$, use

$$PI_{1-\alpha}[Y_{h(\text{new})}] = \left[\hat{Y}_h - \sqrt{g \cdot f_{1-\alpha; g, n-p}} \cdot s_{\hat{Y}_{h(\text{new})}}, \hat{Y}_h + \sqrt{g \cdot f_{1-\alpha; g, n-p}} \cdot s_{\hat{Y}_{h(\text{new})}} \right], \quad (12.87)$$

or

$$\left[\hat{Y}_h - t_{1-\alpha/2g; n-p} \cdot s_{\hat{Y}_{h(\text{new})}}, \hat{Y}_h + t_{1-\alpha/2g; n-p} \cdot s_{\hat{Y}_{h(\text{new})}} \right] \quad (12.88)$$

whichever produces narrower intervals.

12.16.3 Distinguishing Pointwise Confidence Envelopes from Confidence Bands

There is a distinction between connected intervals with $(1 - \alpha) \cdot 100\%$ confidence at each single point and an entire band with $(1 - \alpha) \cdot 100\%$ confidence of containing the regression line. In this text, when each single interval has $(1 - \alpha) \cdot 100\%$ confidence of containing a mean response ($E(Y_h)$), and the upper and lower endpoints of the intervals are connected over the range of possible \mathbf{X}_h values, a **pointwise confidence envelope** is created. If the confidence for containing the entire regression line ($E(Y_h|\mathbf{X}_h)$) is $(1 - \alpha) \cdot 100\%$, a **confidence band** is being calculated. A confidence band is constructed by plotting the upper and lower values of (12.85) over the range of \mathbf{X}_h and subsequently connecting all of the upper values with a curve and all of the lower values with a curve from (12.85). A graphical representation of 90% pointwise confidence intervals, 90% confidence bands, and 90% pointwise prediction intervals for the regression of **gpa** on **sat** using the **Grades** data frame from the **PASWR** package is shown in Figure 12.26 on the following page.

Example 12.26 Use the data **HSwrestler** and the linear model in (12.6) with **HWFAT** as the response and **AGE**, **ABS**, and **TRICEPS** as the predictors, assuming the errors from this model are normally distributed with mean zero and constant variance σ^2 .

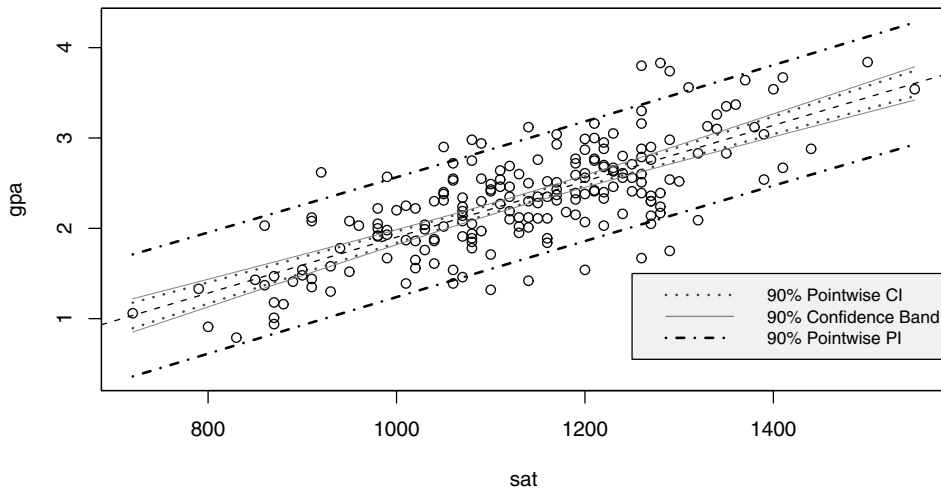


FIGURE 12.26: Representation of 90% pointwise confidence intervals, 90% prediction intervals, and a 90% confidence band for the regression of `gpa` on `sat` using the data in `Grades`

- Obtain joint interval estimates for β_1 , β_2 , and β_3 using a 90% family confidence coefficient with both the Bonferroni and Scheffé approaches.
- Use the function `confidence.ellipse()` from the package `car` to construct a 90% simultaneous confidence region for β_2 and β_3 . Use the function `abline()` to verify visually that the limits of the simultaneous confidence region drawn by `confidence.ellipse()` agree with the values found in part (a).
- Find 90% joint interval estimates for the mean HWFAT of wrestlers with values of \mathbf{X}_{hi} given in Table 12.9.

Table 12.9: Values of \mathbf{X}_{hi} for `HSwrestler`

	<i>AGE</i>	<i>ABS</i>	<i>TRICEPS</i>
\mathbf{X}_{h1}	16	10	9
\mathbf{X}_{h2}	17	11	11
\mathbf{X}_{h3}	18	8	8

- Find 90% joint prediction intervals for three new wrestlers with values of \mathbf{X}_{hi} given in Table 12.9.

Solution: The answers are as follows:

- The estimates of the $\hat{\beta}_k$ s, $s_{\hat{\beta}}$ s, the Bonferroni critical value $t_{1-\alpha/2*g}$, and the Scheffé

critical value $\sqrt{pf_{1-\alpha; q, n-p}}$ are computed with S and then used in (12.82) and (12.84) to compute three simultaneous confidence intervals, respectively.

The Bonferroni simultaneous confidence intervals are

$$CI_{0.90}(\beta_1) = [-1.098, 0.032]$$

$$CI_{0.90}(\beta_2) = [0.219, 0.494]$$

$$CI_{0.90}(\beta_3) = [0.251, 0.680]$$

The Scheffé simultaneous confidence intervals are

$$CI_{0.90}(\beta_1) = [-1.197, 0.130]$$

$$CI_{0.90}(\beta_2) = [0.195, 0.518]$$

$$CI_{0.90}(\beta_3) = [0.214, 0.718]$$

```
> attach(HSwrestler)
> alpha <- 0.10
> mult.model <- lm(HWFAT~AGE+ABS+TRICEPS)
> summary(mult.model)$coef
      Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 10.6160623  4.23272425  2.508092 1.433001e-02
AGE          -0.5330948  0.26067474 -2.045057 4.440545e-02
ABS           0.3564311  0.06353588  5.609918 3.323075e-07
TRICEPS       0.4656071  0.09898493  4.703819 1.158514e-05
> b <- summary(mult.model)$coef[2:4,1]
> s.b <- summary(mult.model)$coef[2:4,2]
> g <- 3
> B <- qt((1-alpha)/(2*g),78-4)
> B
[1] 2.168523
> BonSimCI.b <- matrix(c(b-B*s.b, b+B*s.b), ncol=2)
> conf <- c("5%", "95%")
> bnam <- c("AGE", "ABS", "TRICEPS")
> dimnames(BonSimCI.b) <- list(bnam, conf)
> BonSimCI.b
      5%      95%
AGE    -1.0983739  0.0321843
ABS     0.2186521  0.4942101
TRICEPS 0.2509561  0.6802582

> Q <- 3
> S <- sqrt(Q*qt(.9, Q,78-4))
> S
[1] 2.545185
> SchSimCI.b <- matrix(c(b-S*s.b, b+S*s.b), ncol=2)
> dimnames(SchSimCI.b) <- list(bnam, conf)
> SchSimCI.b
      5%      95%
AGE    -1.1965602  0.1303706
ABS     0.1947205  0.5181416
TRICEPS 0.2136722  0.7175421
```


(b) The following code is used to create the left graph in Figure 12.27, which depicts a joint confidence region for β_2 and β_3 enclosed by the Bonferroni confidence limits:

```
> confidence.ellipse(lm(HWFAT~AGE+ABS+TRICEPS), level=.90,
+ which.coef=c(3,4), Scheffe=FALSE, main="")
> title(main="Bonferroni Confidence Region")
> abline(v=BonSimCI.b[2,])
> abline(h=BonSimCI.b[3,])
```

In a similar fashion, the right graph of Figure 12.27 depicts a joint confidence region for β_2 and β_3 enclosed by the Scheffé confidence limits. The code to reproduce the right graph of Figure 12.27 is

```
> confidence.ellipse(lm(HWFAT~AGE+ABS+TRICEPS), level=.90,
+ which.coef=c(3,4), Scheffe=TRUE, main="")
> title(main="Scheffe Confidence Region")
> abline(v=SchSimCI.b[2,])
> abline(h=SchSimCI.b[3,])
```

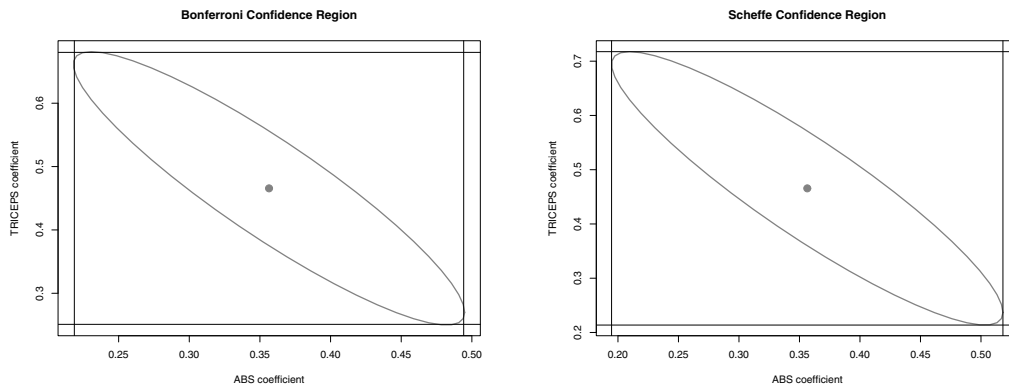


FIGURE 12.27: Joint confidence region for β_2 and β_3 enclosed by the Bonferroni (left graph) and Scheffé (right graph) confidence limits

(c) The 90% simultaneous confidence intervals for the mean HWFAT of wrestlers with values of \mathbf{X}_{hi} given in Table 12.9 on page 644 using (12.86) since $t_{1-\frac{\alpha}{2g}; n-p} = 2.16 < \sqrt{g \cdot f_{1-\alpha; g, n-p}} = 2.55$ are

$$CI_{0.90}[E(Y_{h1})] = [8.998, 10.685]$$

$$CI_{0.90}[E(Y_{h2})] = [9.448, 11.744]$$

$$CI_{0.90}[E(Y_{h3})] = [6.048, 9.145]$$

```
> g <- 3
> alpha <- 0.10
> SC <- sqrt(g*qf(1-alpha,3,74))
```

```

> TC <- qt(1-alpha/(2*g),74)
> c(TC, SC)
[1] 2.168523 2.545185
> RES <- predict(mult.model, newdata=data.frame(AGE=c(16,17,18),
+ ABS=c(10,11,8), TRICEPS=c(9,11,8)), se.fit=TRUE)
> Yhath <- RES$fit
> Syhath <- RES$se.fit
> ll <- Yhath - TC*Syhath
> ul <- Yhath + TC*Syhath
> BCI <- cbind(Yhath, Syhath, ll, ul)
> BCI
      Yhath   Syhath      ll      ul
1  9.841321 0.3888869 8.998010 10.684631
2 10.595871 0.5294386 9.447771 11.743971
3  7.596662 0.7141915 6.047921  9.145402
> round(BCI,3)
      Yhath Syhath      ll      ul
1  9.841  0.389 8.998 10.685
2 10.596  0.529 9.448 11.744
3  7.597  0.714 6.048  9.145

```

(d) The 90% joint prediction intervals for three new wrestlers with values of \mathbf{X}_{hi} given in Table 12.9 on page 644 using (12.81) since $t_{1-\alpha/2g; n-p} = 2.17 < \sqrt{gf_{1-\alpha; g, n-p}} = 2.55$ are

$$PI_{0.90}[Y_{h1(new)}] = [3.285, 16.397]$$

$$PI_{0.90}[Y_{h2(new)}] = [3.994, 17.198]$$

$$PI_{0.90}[Y_{h3(new)}] = [0.913, 14.280]$$

```

> g <- 3
> alpha <- 0.10
> SC <- sqrt(g*qt(1-alpha,3,74))
> TC <- qt(1-alpha/(2*g),74)
> c(SC, TC)
[1] 2.545185 2.168523
> # Use TC with equation 12.69
> MSE <- anova(mult.model)[4,3]
> MSE
[1] 8.989042
> s2yhathnew <- MSE + Syhath^2
> Syhathnew <- sqrt(s2yhathnew)
> ll <- Yhath - TC*Syhathnew
> ul <- Yhath + TC*Syhathnew
> SPI <- cbind(Yhath, Syhathnew, ll, ul)
> SPI
      Yhath Syhathnew      ll      ul
1  9.841321  3.023289 3.2852500 16.39739
2 10.595871  3.044560 3.9936729 17.19807
3  7.596662  3.082063 0.9131382 14.28019
> detach(HSwrestler)

```



12.17 Problems

1. The manager of a URL commercial address is interested in predicting the number of megabytes downloaded, `megasd`, by clients according to the number of minutes they are connected, `mconnected`. The manager randomly selects (megabyte, minute) pairs, records the data, and stores the pairs (`megasd`, `mconnected`) in the file `URLaddress`.
 - (a) Create a scatterplot of the data. Is the relationship between `megasd` and `mconnected` linear?
 - (b) Fit a regression line to the data, and superimpose the resulting line in the plot created in part (a).
 - (c) Compute the covariance matrix of the $\hat{\beta}$ s.
 - (d) What is the standard error of $\hat{\beta}_1$?
 - (e) What is the covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$?
 - (f) Construct a 95% confidence interval for the slope of the regression line.
 - (g) Compute R^2 , R_a^2 , and the residual variance for the fitted regression.
 - (h) Is the normality assumption satisfied?
 - (i) Are there any outlying observations?
 - (j) Are there any influential observations? Compute and graph Cook's distances, DF-FITS, and DFBETAS to answer this question.
 - (k) Estimate the mean value of megabytes downloaded by clients spending 10, 30, and 50 minutes on line. Construct the corresponding 95% confidence intervals.
 - (l) Predict the megabytes downloaded by a client spending 30 minutes on line. Construct the corresponding 95% prediction interval.
2. A metallurgic company is investigating lost revenue due to worker illness. It is interested in creating a table of lost revenue to be used for future budgets and company forecasting plans. The data are stored in the data frame `LostR`.
 - (a) Create a scatterplot of lost revenue versus number of ill workers. Is the relationship linear?
 - (b) Fit a regression line to the data, and superimpose the resulting line in the plot created in part (a).
 - (c) Compute the covariance matrix of the $\hat{\beta}$ s.
 - (d) Create a 95% confidence interval for β_1 .
 - (e) Compute the coefficient of determination and the adjusted coefficient of determination. Provide contextual interpretations of both values.
 - (f) What assumptions need to be satisfied in order to use the model from part (b) for inferential purposes?
 - (g) Create a table of expected lost revenues when 5, 15, and 20 workers are absent due to illness.
 - (h) Compute a 95% prediction interval of lost revenues when 13 workers are absent due to illness.

3. To obtain a linear relationship between the employment (number of employed people = dependent variable) and the GDP (gross domestic product = response variable), a researcher has taken data from 12 regions. Use the following information to answer the questions:

$$\sum_{i=1}^{12} x_i = 581 \quad \sum_{i=1}^{12} x_i^2 = 28507 \quad \sum_{i=1}^{12} x_i Y_i = 2630 \quad \sum_{i=1}^{12} Y_i = 53 \quad \sum_{i=1}^{12} Y_i^2 = 267$$

Source	df	SS	MS	F_{obs}	ϕ -value
Regression	*	*	*	*	*
Error	*	22.08	*	*	*

- Complete the ANOVA table.
 - Decide if the regression is statistically significant.
 - Compute and interpret the coefficient of determination.
 - Calculate the model's residual variance.
 - Write out the fitted regression line and construct a 90% confidence interval for the slope.
4. The speed of a tennis ball after being struck with a tennis racket depends on the length of the racket and the string tension. A multiple regression model is fit where Y is the speed of the struck tennis ball, x_1 is the length of the racket, and x_2 is the string tension, for 16 different tennis rackets. The following table displays the analysis of variance for the fitted regression model:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	--	3797.0	--	--	--
x2	--	1331.3	--	--	--
Residuals	--	174.1	--	--	--

- Complete the table.
- Compute the regression sum of squares. Is the regression statistically significant?
- Estimate the model's error variance.
- Compute both R^2 and R_a^2 coefficients.
- Given $\hat{\beta}_0 = -8.355$, $\hat{\beta}_1 = 3.243$, $\hat{\beta}_2 = -1.711$, $Var[\hat{\beta}_0] = 292.280$, $Var[\hat{\beta}_1] = 0.051$ and $Var[\hat{\beta}_2] = 0.029$, conduct the following tests of hypotheses and comment on the results:

$$\begin{array}{lll} H_0 : \beta_0 = 0 & H_0 : \beta_1 = 3 & H_0 : \beta_2 = -1 \\ H_1 : \beta_0 \neq 0, & H_1 : \beta_1 > 3, & H_1 : \beta_2 < -1. \end{array}$$

5. Given a simple linear regression model, show

- $\hat{\sigma}^2 = \frac{\sum_i \hat{\epsilon}_i^2}{n-2}$ is an unbiased estimator of σ^2 .

(b) The diagonal element of the hat matrix can be expressed as

$$h_{ii} = \mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_i(x_i - \bar{x})^2},$$

where $\mathbf{x}'_i = (1, x_i)$.

6. Show that (12.64) and (12.65) are algebraically equivalent:

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})'(\mathbf{X}'\mathbf{X})(\hat{\beta}_{(i)} - \hat{\beta})}{p\hat{\sigma}^2} = \frac{r_i^2}{p} \frac{h_{ii}}{1 - h_{ii}}.$$

7. Show that (12.66) and (12.67) are algebraically equivalent:

$$\text{DFFITs}_i = \frac{|\hat{Y}_i - \hat{Y}_i(i)|}{\sigma_{(i)}\sqrt{h_{ii}}} = r_i^* \sqrt{\frac{h_{ii}}{1 - h_{ii}}}$$

8. Show that the *SSE* in a linear model expressed in summation notation is equivalent to the *SSE* expressed in matrix notation:

$$\text{SSE} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \mathbf{Y}'\mathbf{Y} - \hat{\beta}\mathbf{X}'\mathbf{Y}$$

9. Show that the *SSR* in a linear model expressed in summation notation is equivalent to the *SSR* expressed in matrix notation:

$$\text{SSR} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = \hat{\beta}'\mathbf{X}'\mathbf{Y} - \frac{1}{n}\mathbf{Y}'\mathbf{J}\mathbf{Y}$$

10. Show that the trace of the hat matrix \mathbf{H} is equal to p , the number of parameters (β s), in a multiple linear regression model.

11. Suppose a realtor wants to model the appraised price of an apartment in Vitoria as a function of the predictors living area and the status of the apartment's conservation. Consider the data frame `vit2005`, which contains data about apartments in Vitoria, Spain, including total price, area, and conservation. The variable conservation has four levels: 1A, 2A, 2B, and 3A.

- Define a new conservation variable called `conservation1` with three levels, *A*, *B*, and *C*, where $A = 1A$, $B = 2A$, and $C = 2B$ and $3A$ together. Define the corresponding dummy variables considering *A* (the first category) as the reference category.
- Write and fit separate linear regression models (different intercepts and different slopes) for each `conservation1` category.
- Construct a single scatterplot of the data where the fitted models are superimposed over the scatterplot.

Case Study: Biomass

Data for this case study come from *Gobierno de Navarra* and *Gestión Ambiental de Viveros y Repoblaciones de Navarra*, 2006.

The data were obtained within the European Project FORSEE.

12. To estimate the amount of carbon dioxide retained in a tree, its biomass needs to be known and multiplied by an expansion factor (there are several alternatives in the literature). To calculate the biomass, specific regression equations by species are frequently used. These regression equations, called allometric equations, estimate the biomass of the tree by means of some known characteristics, typically diameter and/or height of the stem and branches. The **biomass** file contains data of 42 beeches (*Fagus Sylvatica*) from a forest of Navarra (Spain) in 2006, where

- **Dn**: diameter of the stem in centimeters
 - **H**: height of the tree in meters
 - **PST**: weight of the stem in kilograms
 - **PSA**: aboveground weight in kilograms
- (a) Create a scatterplot of **PSA** versus **Dn**. Is the relationship linear? Superimpose a regression line over the plot just created.
 - (b) Create a scatterplot of $\log(\text{PSA})$ versus $\log(\text{Dn})$. Is the relationship linear? Superimpose a regression line over the plot just created.
 - (c) Fit the regression model $\log(\text{PSA}) = \beta_0 + \beta_1 \log(\text{Dn})$, and compute R^2 , R_a^2 , and the variance of the residuals.
 - (d) Introduce **H** as an explanatory variable and fit the model $\log(\text{PSA}) = \beta_0 + \beta_1 \log(\text{Dn}) + \beta_2 \text{H}$. What is the effect of introducing **H** in the model?
 - (e) Complete the ANALYSIS QUESTIONS for the model in (d).

ANALYSIS QUESTIONS:

- (1) Estimate the model's parameters and their standard errors. Provide an interpretation for the model's parameters.
 - (2) Compute the variance-covariance matrix of the $\hat{\beta}$ s.
 - (3) Provide 95% confidence intervals for $\hat{\beta}_1$ and $\hat{\beta}_2$.
 - (4) Compute the R^2 , R_a^2 , and the residual variance.
 - (5) Construct a graph with the default diagnostics plots of R.
 - (6) Can homogeneity of variance be assumed?
 - (7) Do the residuals appear to follow a normal distribution?
 - (8) Are there any outliers in the data?
 - (9) Are there any influential observations in the data?
- (f) Obtain predictions of the aboveground biomass of trees with diameters **Dn** = `seq(12.5, 42.5, 5)` and heights **H** = `seq(10, 40, 5)`. Note that the weight predictions are obtained from back transforming the logarithm. The bias correction is obtained by means of the lognormal distribution: If \hat{Y}_{pred} is the prediction, the corrected (back-transformed) prediction \tilde{Y}_{pred} is given by

$$\tilde{Y}_{\text{pred}} = \exp(\hat{Y}_{\text{pred}} + \hat{\sigma}^2/2)$$

where $\hat{\sigma}^2$ is the variance of the error term.

- (g) Fit the following regression model for the weight of the stem. $\text{PST} = \beta_0 + \beta_1 \text{Dn} + \beta_2 \text{H}$.
- (h) Display the default diagnostics plots. What does the plot of the fitted values versus the residuals suggest?

- (i) Propose a model to correct the problem from part (h).
- (j) Does your new model correct the residuals problem detected in (h)?

Case Study: Fruit Trees

Data and ideas for this case study come from Militino et al. (2006).

13. To estimate the total surface occupied by fruit trees in three small areas (R63, R67, and R68) of Navarra in 2001, a sample of 47 square segments has been taken. The experimental units are square segments or quadrats of 4 hectares, obtained by random sampling after overlaying a square grid on the study domain. The focus of this case study is to illustrate two different techniques used to obtain estimates: direct estimation and small area estimation. The direct technique estimates the total surface area by multiplying the mean of the occupied surface in the sampled segments by the total number of segments in every small area. The small area technique consists of estimating a regression model where the dependent variable is the observed surface area occupied by fruit trees in every segment and the explanatory variables are the classified cultivars by satellite in the same segment and the small areas where they belong to. The final surface area totals are obtained by multiplying the total classified surface area of every small area by the β 's parameter estimates obtained from the regression model (observed surface area \sim classified surface area + small areas).

The surface variables in the data frame `satfruit` are given in m^2 :

- QUADRAT is the number of sampled segment or quadrat
 - SArea are the small areas' labels
 - WH is the classified surface of wheat in the sampled segment
 - BA is the classified surface of barley in the sampled segment
 - NAR is the classified surface of fallow or non-arable land in the sampled segment
 - COR is the classified surface of corn in the sampled segment
 - SF is the classified surface of sunflower in the sampled segment
 - VI is the classified surface of vineyard in the sampled segment
 - PS is the classified surface of grass in the sampled segment
 - ES is the classified surface of asparagus in the sampled segment
 - AF is the classified surface of lucerne in the sampled segment
 - CO is the classified surface of rape *Brassica napus* in the sampled segment
 - AR is the classified surface of rice in the sampled segment
 - AL is the classified surface of almonds in the sampled segment
 - OL is the classified surface of olives in the sampled segment
 - FR is the classified surface of fruit trees in the sampled segment
 - OBS is the observed surface of fruit trees in the sampled segment
- (a) Characterize the shape, center, and spread for the variable FR.
 - (b) What is the maximum number of m^2 of classified fruits by segment?
 - (c) How many observations are there by small area?
 - (d) Use `pairs()` to explore the linear relationships between OBS and the remainder of the numerical variables. Comment on the results.

- (e) Create histograms of the observed fruits surface area (**OBS**) by small areas (**SArea**).
- (f) Use boxplots and barplots with standard errors to compare the observed surface area (**OBS**) and the classified surface area (**FR**) by small areas (**SArea**).
- (g) Compute the correlation between **OBS** and all other numerical variables. List the three variables in order along with their correlation coefficients that have the highest correlation with **OBS**.

Model (A) Fit the linear regression model, called Model (A), of **OBS** versus the rest of the numerical variables in the same order as they are recorded in the file.

- i. Do an ANOVA and decide which variables are statistically significant. Use $\alpha = 0.05$.
- ii. Compute the coefficient of determination R^2 , R_a^2 , the AIC, and the BIC statistic. What is the proportion of total variability explained by Model (A)?

Model (B) Find the best regression model using `leaps()` with the R_a^2 method, from the package `leaps`. Call this Model (B).

Model (C) Find the best regression model using `step()` to determine the best subset regression. Call this Model (C).

- i. Check that the coefficient of determination R^2 , the adjusted R_a^2 , the AIC, and the BIC of Models (A), (B), and (C) are the following:

Model	R^2	R_a^2	AIC	BIC
Model (A)	0.78	0.69	880	909
Model (B)	0.78	0.72	871	891
Model (C)	0.75	0.72	867	878

What is the best model using both AIC and BIC statistics? Why?

- ii. Check that `leaps()` chooses the variables **SF**, **PS**, **ES**, **AF**, **CO**, **AR**, **AL**, **OL**, and **FR** and `step()` chooses **PS**, **AL**, **OL**, and **FR**.
- iii. Graph the default diagnostic regression plots of Model (C). Plot the standardized residuals, the student residuals, the Cook distances, the diagonal elements of the hat matrix, the DFFITS, and DFBETAS of Model (C).
- iv. Are there any leverage points?
- v. Are there any outliers?
- vi. Test the normality hypothesis with `shapiro.test`. Check graphically the absence of heteroscedasticity in Model (C).

Model (D) Introduce **SArea** in Model (A). Choose the best model using `step` and call it Model (D).

- i. Do an ANOVA for Model (D). What variables are statistically significant? Calculate 95% confidence intervals for the β s of the explanatory variables.
- ii. Check that the coefficient of determinations R^2 , R_a^2 , the AIC, and the BIC statistic of Model (D) are the following:

Model	R^2	R_a^2	AIC	BIC
Model (D)	0.81	0.79	855	868

- iii. Use the function `drop1()` to test the statistically significant presence of **PS** and **AL**.
- iv. Use the `confidence.ellipse()` function from the package `car` to test that **PS** and **AL** are jointly equal to zero.

Model (E) Drop out the variables PS and AL of Model (D). The new model is called Model (E).

- i. Do the default diagnostic regression plots of Model (E).
- ii. Plot the standardized residuals, the student residuals, the Cook distances, the diagonal elements of the hat matrix, the DFFITS, and DFBETAS of Model (E).
- iii. Are there any leverage points? Justify the answer given.
- iv. Are there any outliers? Justify the answer given.
- v. Check normality and homoscedasticity for Model (E) using graphics and hypotheses tests.

Model (F) Drop out the 46 record of Model (E). Fit the new model and call it Model (F).

- i. Do the default diagnostic regression plots of Model (F).
 - ii. Plot the standardized residuals, the student residuals, the Cook distances, the diagonal elements of the hat matrix, the DFFITS, and DFBETAS of Model (F).
 - iii. Are there any leverage points?
 - iv. Are there any outliers?
 - v. Check the adequacy of the normality and homoscedasticity assumptions of Model (F).
 - vi. Compute 95% confidence intervals for the parameters of the explanatory variables in Model (F) and comment on the results.
- (h) How many hectares of observed fruits are expected to be incremented if the classified hectares of fruit trees by the satellite are increased by 10000 m² (1 ha)?
- (i) Suppose the total classified fruits by the satellite in area R63 is 97044.28 m², in area R67 is 4878603.43 m², and in area R68 is 2883488.24 m². Calculate the total prediction of fruit trees by small areas.
- (j) Plot in the same graphical page FR versus OBS separately by the three areas. Superimpose the corresponding regression lines.
- (k) Plot the individual predictions versus the observed data. Add a diagonal line to the plot.
- (l) Do a barplot to graph simultaneously the predicted totals and the direct estimates by areas knowing that the total number of classified segments in areas R63, R67, and R68, are 119, 703, and 564, respectively..

Case Study: Real Estate

Data from Departamento de Economía y Hacienda de la Diputación Foral de Álava, and LKS Tasaciones

14. The goal of this case study is to walk the user through the creation of a parsimonious multiple linear regression model that can be used to predict the total price (`totalprice`) of apartments by their hedonic (structural) characteristics. The data frame `vit2005` contains several variables, and further description of the data can be found in the help file.
- (a) Plot `totalprice` versus the numerical explanatory variables `area`, `age`, `floor`, `rooms`, `toilets`, `garage`, `elevator`, and `tras` to see if these variables have a linear relationship with `totalprice`.

Model (A) Create a linear regression model, called Model (A), between `totalprice` as the response variable and the rest of the variables as explanatory variables in the same order as they appear in the file.

- (i) Do an analysis of variance of Model (A). Decide which variables are statistically significant using $\alpha = 0.05$.
- (ii) Verify that if `age` and `floor` are specified last in the model, then `age` is not statistically significant.

Model (B) Load the `leaps` package. Using `leaps()` with the R_a^2 method, determine the best regression subset, and call this subset Model (B).

- (i) The function `leaps()` excludes `conservation` but not the `age` variable. This happens because both variables are correlated. Create a boxplot of age for each level of conservation.
- (ii) Does the boxplot help to explain the correlation between age and conservation?

Model (C) Use the `step()` command and determine the best regression subset. Call this Model (C).

- (i) Comment on the results of Model (C)
- (ii) Compare the results with those obtained from Model (B). In other words, do the procedures `step()` and `leaps()` select the same variables?

Model (D) Define a new model using the name Model (D) with the intersection of the variables from models (B) and (C).

- (i) Find R^2 , R_a^2 , AIC, and BIC for Model (D).
 - (ii) Compare R^2 , R_a^2 , AIC, and BIC for Model (D) with those values as obtained from Models (A), (B), and (C).
- (b) Graph the default diagnostic regression plots of Models (A), (B), (C), and (D), and test for the models' assumptions.
 - (c) Load the `MASS` package. Choose a transformation of the Box-Cox family to reduce the heteroskedasticity in Model (D).
 - (d) Compute the correlations between `log(totalprice)` and the rest of the quantitative variables in `vit2005`. Compare the results with those obtained with `totalprice`.
 - (e) Plot `log(totalprice)` versus the numerical explanatory variables `area`, `age`, `floor`, `rooms`, `toilets`, `garage`, `elevator`, and `tras` one at a time. Is the relationship between `log(totalprice)` and the chosen explanatory variables linear?

Model (E) Find the linear regression model between `log(totalprice)` and the rest of the variables in data frame `vit2005`.

- (i) Do the analysis of variance for Model (E).
- (ii) What variables are statistically significant?
- (iii) Is it possible to select an appropriate model using the analysis of variance?

Model (F) Let Model (F) be the model that results when the R_a^2 criterion and the function `leaps()` are used to select a model.

Model (G) Let Model (G) be the model that results from using the function `step()` to select a model.

- (i) Find R^2 , R^2 , R_a^2 , AIC, and BIC for Models (E), (F), and (G), and interpret the results.
- (ii) Compute the model's parameters of Model (G) and their standard errors.

- (iii) Find the variance inflation factors for Model (G). Is multicollinearity a problem?
 - (iv) Graph the diagnostic regression plots of Model (G). Check if the model assumptions are valid.
 - (v) Construct and interpret 95% confidence intervals for the parameters in Model (G).
 - (vi) Check graphically the linearity of the explanatory variables `area` and `age` in Model (G). Create a graph of the standardized residuals versus the explanatory variables.
 - (vii) Plot the standardized residuals, the studentized residuals, the Cook distances, the diagonal elements of the hat matrix, the DFFITS, and the DFBETAS (variables `area` and `age`) of Model (G).
 - (viii) Are there any leverage points?
 - (ix) Are there any outliers?
- (f) Drop separately and jointly both observations 3 and 93 and refit Model (G).
- Model (H) Drop observation 3 to obtain Model (H).
- Model (I) Drop observation 93 to obtain Model (I).
- Model (J) Drop both observations 3 and 93 to obtain Model (J).
- (i) Find R^2 , R_a^2 , AIC, and BIC for Models (H), (I), and (J). Comment on the results.
 - (ii) Graph the model regression diagnostics of Models (H), (I), and (J). Check if the models' assumptions are satisfied.
- (g) Use `drop1()` to check whether or not every explanatory variable in Model (J) is statistically significant.
- Model (K) Define a new Model (K) without the `streetcategory` variable.
- (i) Find R^2 , R_a^2 , AIC, and BIC of Model (K). Compare them with those obtained for Model (J).
 - (ii) Graph the default diagnostic regression plots of Model (K), and test the model's assumptions.
 - (iii) Calculate the influence measures for Model (K).
 - (iv) Are there any leverage points? Justify.
 - (v) Are there any outliers? Justify.
 - (vi) Find the parameter estimates, and compute 95% confidence intervals for the parameters of Model (K).
 - (vii) Find the relative contribution of the explanatory variables to explaining the variability of the prices in Model (K).
 - (viii) What is the variable that explains the most variability?
 - (ix) What variables jointly explain 80% of the total variability of `log(totalprice)`?
 - (x) Find the predictions of Model (K) (a) with bias correction and (b) without bias correction. The bias correction is obtained by means of the lognormal distribution: If \hat{Y}_{pred} is the prediction of Model (K), the corrected (backtransformed) prediction \tilde{Y}_{pred} of Model (K) is given by

$$\tilde{Y}_{\text{pred}} = \exp(\hat{Y}_{\text{pred}} + \hat{\sigma}^2/2)$$

where $\hat{\sigma}^2$ is the variance of the error term, and the confidence interval is given by

$$l_{\text{inf}} = \exp(\widehat{Y}_{\text{pred}} + \widehat{\sigma}^2/2 - z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}(\widehat{Y}_{\text{pred}}) + \widehat{\text{Var}}(\widehat{\sigma}^2)/4})$$

$$l_{\text{sup}} = \exp(\widehat{Y}_{\text{pred}} + \widehat{\sigma}^2/2 + z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}(\widehat{Y}_{\text{pred}}) + \widehat{\text{Var}}(\widehat{\sigma}^2)/4})$$

$$\text{and } \widehat{\text{Var}}(\widehat{\sigma}^2) = \frac{2\widehat{\sigma}^4}{df_{\text{residual}}}$$

- (xi) For Model (K), plot the predicted values (with and without bias correction) versus observed values. Comment on the results.
- (xii) Show that in Model (K) an increment of 10 m² in the area of a flat implies an increment of 4% in the predicted total price. To verify this, find the predicted price of three apartments with `areas` 80, 90, and 100 m², respectively, and keep the rest of the explanatory variables fixed. For example, assign them the following values: `category=2A`, `age=10`, `toilets=1`, `garage=1`, `elevator=1`, `out=50`, `rooms=3`, `zone=41`, and `tras=1`. Compute the corresponding 95% prediction intervals.
- (xiii) What is the percentage change in the total price of an apartment when the number of garages changes from one to two?
- (xiv) What is the percentage change in the total price of an apartment when the heating type changes from “1A” to “3B”?

Appendix A

S Commands

Table A.1: Useful Commands When Working with Numeric Vectors

Function	Description
<code>cbind(x,y)</code>	Joins vectors x and y as columns of vectors
<code>cor(x,y)</code>	Computes the correlation coefficient
<code>cos(x)</code>	Returns the cosine for all values in x
<code>exp(x)</code>	Computes e^x for all values in x
<code>fivenum(x)</code>	Computes the smallest value, the lower hinge, the median, the upper hinge, and the largest value of a vector x in R
<code>floor(x)</code>	Returns a numeric vector containing the largest integers not greater than the corresponding elements of x
<code>IQR(x)</code>	Returns the interquartile range of x
<code>length(x)</code>	The number of values in x
<code>log(x)</code>	Computes the natural logarithm for all values in x
<code>log10(x)</code>	Computes the base 10 logarithm for all values in x
<code>mad(x,constant=1)</code>	Returns the median absolute deviation of x
<code>max(x)</code>	The largest value of x
<code>mean(x)</code>	Computes the sample mean of x
<code>median(x)</code>	Returns the sample median of x
<code>min(x)</code>	The smallest value of x
<code>prod(x)</code>	The product of all the values in x
<code>quantile(x)</code>	Computes the quantiles of a data set stored in a vector x
<code>range(x)</code>	Returns the smallest and largest values in x
<code>rbind(x,y)</code>	Joins vectors x and y as rows of vectors
<code>rep(x,n)</code>	Repeats vector x n times
<code>round(x,n)</code>	Rounds the number of decimals to n for object x
<code>scale(x)</code>	Computes the z -score of x
<code>sd(x)</code>	Computes the sample standard deviation of x in R
<code>seq(x,y,n)</code>	Creates a sequence of numbers from x to y with incremental value n
<code>sin(x)</code>	Returns the sine for all values in x

Table A.1: Useful Commands When Working with Numeric Vectors (continued)

Function	Description
<code>sqrt(x)</code>	Computes the square root for all values in x
<code>stdev(x)</code>	Computes the sample standard deviation of x in S-PLUS
<code>sum(x)</code>	The sum of all the values in x
<code>summary(x)</code>	Returns the minimum, the first quartile, the median, the mean, the third quartile, and the maximum of x
<code>tan(x)</code>	Returns the tangent for all values in x
<code>var(x)</code>	Computes the sample variance of x
<code>which(x==n)</code>	Give the index of number n in vector x

Table A.2: S Vector and Matrix Functions

Function	Description
<code>A%*%B</code>	Matrix multiplication of A and B
<code>diag(matrix)</code>	Extracts the diagonal elements of the matrix
<code>diag(vector)</code>	Produces a diagonal matrix with the elements from the vector
<code>dim(matrix)</code>	Obtains the matrix dimensions
<code>dimnames(matrix)</code>	Verifies if rows and columns names have been assigned
<code>eigen(matrix)</code>	Used to compute eigenvalues and eigenvectors
<code>matrix(vector, nrow=r, byrow=TRUE)</code>	Creates a matrix by rows with n rows
<code>names(vector)</code>	Allows the assignment of names to a vector
<code>order(vector, matrix, data frames)</code>	Orders by more than one variable
<code>set.seed(number)</code>	Reproduces the same set of pseudo-random numbers
<code>solve(A)</code>	Used to find the inverse of a matrix \mathbf{A}
<code>solve(A,b)</code>	Used to solve systems of equations $\mathbf{Ax} = \mathbf{b}$
<code>sort(vector)</code>	Produces an ordered vector
<code>svd(matrix)</code>	Used to find the singular value decomposition of a matrix
<code>t(A)</code>	Used to find the transpose of a matrix \mathbf{A}

Table A.3: S Functions Used with Arrays, Factors, and Lists

Function	Description
<code>aggregate(x, y, FUN, ...)</code>	Applies function <i>FUN</i> to each element of <i>x</i> based on the categories stored in <i>y</i> and gives the results in a data frame
<code>apply(X, MARGIN, FUN, ...)</code>	Applies function <i>FUN</i> to each <i>MARGIN</i> of <i>X</i> , where <i>X</i> is an array. <i>MARGIN</i> is a vector giving the subscripts over which the function will be applied. 1 indicates rows, 2 indicates columns, 'c(1,2)' indicates rows and columns, etc.
<code>attach(object)</code>	Makes the columns of an object (e.g. data frame) available by names
<code>detach(object)</code>	Detaches an object when finished working with it
<code>dump()</code>	Saves the contents of an S object
<code>ftable()</code>	Creates a compact three-way contingency table
<code>head()</code>	Shows the first six rows of a data frame
<code>lapply(X,FUN)</code>	Applies function <i>FUN</i> to each element of <i>X</i> , where <i>X</i> is a list and the answer is given in the form of a list
<code>load()</code>	Reads a file created with <code>save()</code> in R
<code>margin.table()</code>	Adds margins to a contingency table
<code>prop.table()</code>	Calculates proportions in a contingency table
<code>read.table()</code>	Reads a file in table format
<code>row.names()</code>	Can be used to assign character names to rows of a data frame
<code>sapply(X,FUN)</code>	Calls the function <code>lapply</code> , which applies function <i>FUN</i> to each element of <i>X</i> , where <i>X</i> is either a list or vector. Note that even if <i>X</i> is a list, <code>sapply(X,FUN)</code> returns either a vector or matrix, not a list, as does <code>lapply(X,FUN)</code>
<code>save()</code>	Writes an external representation of R objects
<code>scan()</code>	Reads data into a vector or list from the console or file
<code>source()</code>	Reads a dumped file
<code>split(x,f)</code>	Returns a list of vectors containing the values for the resulting groups when the vector <i>x</i> is split by the factor <i>f</i>
<code>table()</code>	Creates a contingency table based on the supplied factors
<code>tail()</code>	Shows the last six rows of a data frame
<code>tapply(x,y,FUN)</code>	Applies function <i>FUN</i> to each element of <i>x</i> based on the categories stored in <i>y</i> .
<code>write.table()</code>	Allows the contents of an S data frame or matrix to be saved to an external file in ASCII format

Table A.4: Important Probability Distributions That Work with `rdist`, `pdist`, `ddist`, and `qdist`

Distribution	S name	Parameters
beta	<code>beta</code>	<code>shape1</code> , <code>shape2</code>
binomial	<code>binom</code>	n , p
chi-square	<code>chisq</code>	$df = \nu$
exponential	<code>exp</code>	λ
F	<code>f</code>	ν_1, ν_2
Gamma	<code>gamma</code>	<code>shape</code> , <code>rate</code>
geometric	<code>geom</code>	p
hypergeometric	<code>hyper</code>	m, n, k , where m = number of black balls in urn n = number of white balls in urn k = number of balls drawn from the urn
negative binomial	<code>nbinom</code>	n , p
normal	<code>norm</code>	μ, σ
Poisson	<code>pois</code>	λ
Student's t	<code>t</code>	$df = \nu$
uniform	<code>unif</code>	a, b
Weibull	<code>weibull</code>	<code>shape</code> , <code>scale</code>
Wilcoxon rank sum	<code>wilcox</code>	n, m (number of observations on the first and second sample, respectively)
Wilcoxon signed rank	<code>signrank</code>	n

Table A.5: Useful Functions in S for Parametric Inference

Function	Description
<code>fisher.test(x,y=NULL, ...)</code>	Performs Fisher's exact test for testing the null hypothesis of independence between rows and columns in a contingency table (x) with fixed marginals.
<code>prop.test(x,n,p, alternative="two.sided", conf.level=0.95, correct=TRUE)</code>	Compares proportions against hypothesized values, where x is a vector of successes, n is vector containing the number of trials, and p is a vector of probabilities of success specified by the null hypothesis. A continuity correction (<code>correct=TRUE</code>) is used by default.
<code>t.test(x,y=NULL, alternative="two.sided", mu=0, paired=FALSE, var.equal=TRUE, conf.level=.95)</code>	Performs a one-sample, two-sample, or paired t -test, or a Welch modified two-sample t -test. In S-PLUS the default value for <code>var.equal</code> is <code>TRUE</code> .
<code>var.test(x,y, ratio=1, alternative="two.sided", conf.level=0.95)</code>	Performs an F -test to compare the variances of two samples from normal populations.

Table A.6: Useful Functions in S for Nonparametric Inference

Function	Description
<code>binom.test(x,n,p=0.5, alternative="two.sided")</code>	Performs an exact test of a simple null hypothesis about the probability of success in a Bernoulli experiment.
<code>chisq.gof(x, n.classes, mean=mean(x), cut.points=NULL, distribution ="normal", n.param.est=0)</code>	Not directly available in R. Performs a chi-square goodness-of-fit test, where x is numeric vector of observations, <code>n.classes</code> specifies the number of cells into which the observations are to be allocated, <code>cut.points</code> is a vector of cutpoints that define the cells, <code>distribution</code> is a character string that specifies the hypothesized distribution, and <code>n.param.est</code> is the number of parameters estimated from the data.
<code>chisq.test(x,y=NULL, correct=TRUE)</code>	Performs a Pearson's chi-square test on a two-dimensional contingency table, where x is either a matrix or a contingency table.
<code>friedman.test(y, groups, blocks)</code>	Performs a Friedman rank-sum test with unrepeated blocked data, where y is numeric vector, <code>groups</code> is a category object specifying group membership, and <code>blocks</code> is a category object specifying the block membership.
<code>kruskal.test(y, groups)</code>	Performs a Kruskal-Wallis rank-sum test on data, where y is a numeric vector and <code>groups</code> denotes a category object of the same length as y , specifying the group for each corresponding element of y .
<code>ks.gof(x, y=NULL, distribution = "normal", alternative="two.sided")</code>	(Not directly available in R; however, <code>ks.test</code> in R provides similar results to <code>ks.gof</code> .) Performs a one- or two-sample Kolmogorov-Smirnov test, which tests the relationship between two distributions.
<code>wilcox.test(x,y, alternative="two.sided", mu=0, paired=FALSE, exact=FALSE, correct=TRUE)</code>	Computes Wilcoxon signed-rank test for paired or one-sample data and Wilcoxon rank-sum test (Mann-Whitney test) for two-sample data.

Table A.7: Useful Functions in S for Linear Regression and Analysis of Variance

Function	Description
<code>aov(formula,data)</code>	Fits an analysis of variance model according to the specified formula using the specified data.
<code>coefficients(lm object)</code>	Returns the coefficients from a fitted linear regression model. A shorter command with identical results is <code>coef(lm object)</code> .
<code>data.matrix(proj(aov.object))</code>	Returns columns containing estimates for both factors and residuals of <code>aov</code> -type objects.
<code>formula(lm object)</code>	Returns the formula used to fit the linear model.
<code>lm(formula,data)</code>	Fits a linear model to the data according to the user-specified formula.
<code>ls.diag(lm object)</code>	A command used on a <code>ls</code> - or <code>lm</code> -type of object that returns a list containing several quantities for assessing the fit of a least squares regression model including the standard deviation of the residuals, studentized residuals, and the standard errors of the parameter estimates.
<code>lsfit(explanatory variables, response variable(s))</code>	Fits a model using least squares multivariate regression. A list of the estimated coefficients and residuals as well as the QR decomposition of the matrix of explanatory variables is returned. Although the fitted model from <code>lsfit()</code> is identical to <code>lm()</code> , the manner in which the model is specified and the output for the two functions are different.
<code>model.matrix(lm object)</code>	Creates a design matrix.
<code>multicomp(aov object)</code>	(Not directly available in R; however, <code>TukeyHSD()</code> provides similar results to <code>multicomp()</code> when the default argument for <code>method</code> , <code>"best.fast,"</code> is changed to <code>method="tukey".</code>) S-PLUS function <code>multicomp()</code> computes simultaneous or non-simultaneous confidence intervals or bounds for the specified estimatable linear combinations of the parameters in a fixed effects model (stored in <code>aov</code>).
<code>shapiro.test(x)</code>	Computes the Shapiro-Wilk W -statistic, a well-known goodness-of-fit test for the normal distribution.

Table A.8: Useful Contrast Functions in S for Linear Regression and Analysis of Variance

Function	Description
<code>contr.helmert()</code>	Returns a matrix of orthogonal contrasts for different combinations of the factor levels. Contrast i is the difference between level $i + 1$ and the average of levels 1 through i .
<code>contr.poly()</code>	Returns a matrix of orthogonal contrasts for different combinations of the factor levels. Creates orthogonal polynomials of degree 1, 2, etc., either on equally spaced points if n was a single number, or on the points specified by x . Columns are scaled to have norm 1.
<code>contr.sum()</code>	Returns a matrix of non-orthogonal contrasts.
<code>contr.treatment()</code>	Returns the coding that is not technically a set of contrasts at all.

Table A.9: Useful Model Building Functions in S for Linear Regression and Analysis of Variance

Function	Description
<code>add1(lm object, ~.+ explanatory.variables)</code>	Returns information on models that have one more term than the given object. Tilde, \sim , is the symbol used by S to separate the response variable from the explanatory variables.
<code>drop1(lm object)</code>	Returns total sum of squares, residual sum of squares, and AIC each time a variable is dropped from a regression model. S-PLUS also reports the C_P each time a variable is dropped from the model.
<code>leaps(explanatory variables, response variable)</code>	Returns C_P , R_{adj}^2 , and R^2 so user can select the best regressions using a subset of the given explanatory variables.
<code>step(lm object, scope, ...)</code>	Performs stepwise model selection. The starting model is specified in the first argument (<code>lm object</code>) and the range of models is specified in the <code>scope</code> argument.
<code>update(lm object, ~.± explanatory.variables)</code>	Allows a linear model object to be updated by including, eliminating, or modifying the variables.

Table A.10: Useful Diagnostic Functions in S for Linear Regression and Analysis of Variance

Function	Description
<code>fitted(lm object)</code>	Returns the fitted values from the fitted linear model.
<code>plot(lm object)</code>	S-PLUS creates six graphs (four in R) showing various residuals graphs and a graph of Cook's distance values for the fitted linear model.
<code>predict(lm object)</code>	Returns a vector or an array of predictions using the model specified in the <code>lm</code> object.
<code>residuals(lm object)</code>	Returns the residuals for the fitted linear model. A shorter command with identical results is <code>resid(lm object)</code> .
<code>summary(lm object)</code>	Returns a complete statistical summary for the fitted linear model.

Table A.11: Trellis Functions

Function	Description
<code>barchart(f~x z)</code>	Bar chart, categorized according to <code>f</code>
<code>bwplot(f~x z)</code>	Boxplots for the levels of <code>f</code> (a factor) conditioning on <code>z</code> (another factor)
<code>densityplot(~x z)</code>	Density estimate graph
<code>dotplot(f~x z)</code>	Dotplot with data categorized by <code>f</code>
<code>histogram(~x z)</code>	Histogram
<code>qq(f~x z)</code>	Quantile-quantile plot, <code>f</code> having two levels
<code>qqmath(~x z)</code>	Quantile-quantile graph of a data set versus a distribution's quantiles
<code>stripplot(f~x z)</code>	Strip plots for the levels of <code>f</code> (a factor) conditioning on <code>z</code> (another factor)
<code>xyplot(y~x z)</code>	<i>x-y</i> scatterplot

Note: `x`, `y`, and `z` represent any numeric variable and `f` any factor or character variable.

Table A.12: Basic Plotting Functions

Function	Description
<code>abline(a, b)</code>	Adds a straight line with intercept <code>a</code> and slope <code>b</code> .
<code>abline(h=c,...)</code>	Draws a horizontal line at $y = c$.
<code>abline(v=c,...)</code>	Draws a vertical line at $x = c$.
<code>identify(x, y, labels)</code>	Reads the position of the graphics pointer when the left mouse button is clicked. <code>x</code> and <code>y</code> are the coordinates of points in a scatterplot and are required arguments. <code>labels</code> is an optional vector, the same length as <code>x</code> and <code>y</code> , giving labels for the points.
<code>legend(x, y, legend, ...)</code>	Adds a legend to the current plot, where <code>x</code> and <code>y</code> determine the legend coordinates, and <code>legend</code> is a vector of text values to appear in the legend. To determine the coordinates where we want to place our legend, use the function <code>locator()</code> . It is possible to combine the functions <code>legend()</code> and <code>locator()</code> into one step by using <code>legend(locator(), legend,...)</code> .
<code>lines(x,y,...)</code>	Adds points or lines to the current plot.
<code>locator()</code>	Reads the position of the graphics cursor when the left mouse button is pressed.
<code>points(x, y,...)</code>	Adds points or lines to the current plot at the coordinates specified in the vectors <code>x</code> and <code>y</code> .
<code>segments(x1,y1,x2,y2)</code>	Adds the line segment AB with coordinates $A = (x_1, y_1)$ and $B = (x_2, y_2)$ to an existing graph.
<code>text(x, y, labels)</code>	Draws the strings given in the vector <code>labels</code> at the coordinates given by <code>x</code> and <code>y</code> . Note: <code>labels</code> is one or more character strings or expressions specifying the text to be written.
<code>title("Title")</code>	Adds titles to the current plot. To create a multi-line title, type <code>\n</code> at each place we want the text to start another line.

Table A.13: Graphs Frequently Used with Descriptive Statistics

Function	Description
<code>barplot(height, ...)</code>	Creates a bar plot with vertical or horizontal bars where <code>height</code> is a matrix or vector giving the heights (positive or negative) of the bars.
<code>boxplot(x)</code>	Produces a boxplot of the values stored in <code>x</code> .
<code>boxplot(split(x, f))</code>	Produces side-by-side boxplots of the values in <code>x</code> based on the factor <code>f</code> .
<code>bwplot()</code>	Produces horizontal boxplots in S-PLUS.
<code>dotchart(formula, ...)</code>	Creates a multi-way dotplot. Note: Arguments for <code>dotchart</code> in R and S-PLUS are different. See respective help systems for details.
<code>hist(x, ...)</code>	Creates a histogram of the values in <code>x</code> .
<code>interaction.plot(x.factor, trace.factor, response, ...)</code>	Plots the mean (or other summary) of the response for two-way combinations of factors where <code>x.factor</code> contains the levels for the x axis, <code>trace.factor</code> is another factor whose levels form the traces, and <code>response</code> is a numeric variable containing the responses at the various factor combinations.
<code>lines(density())</code>	Adds a density to an existing plot (for example, a histogram).
<code>pairs(x,...)</code>	Creates a scatterplot for each pair of variables in <code>x</code> .
<code>persp(x, y, z,...)</code>	Three-dimensional perspective plots. See system help for more details.
<code>pie(x,...)</code>	Produces a pie chart where the values in <code>x</code> are displayed as the areas of the pie slices.
<code>plot(x,...)</code>	Generic function for plotting S objects.
<code>plot(x,y)</code>	Produces a scatterplot
<code>plot.design(x,y, fun=)</code>	Plots univariate effects of one or more factors, typically for a designed experiment. A function such as mean or median must be typed after <code>fun=</code> , which is then applied to each subset.
<code>qqnorm(x)</code>	Produces a quantile-quantile plot that is used to assess how close the values in <code>x</code> follow the normal distribution.
<code>qqline(x)</code>	Plots a line through the first and third quartiles of the data, and the corresponding quantiles of the standard normal distribution.
<code>qqplot(x,y)</code>	Creates a quantile-quantile plot
<code>stem(x)</code>	Creates a stem-and-leaf plot
<code>stripchart(x~g)</code>	Creates a dotplot or strip chart in R that permits one to compare the distribution of x over g groups

Table A.14: Commonly Used Graphical Parameters

Parameter	Description
<code>adj=0</code>	String justification: 0 means left justify, 1 means right justify, .5 means center the text. Other numbers are a corresponding distance between the extremes.
<code>axes=TRUE/axes=FALSE</code>	<code>axes=TRUE</code> draws a box around the graph, which is the default value. <code>axes=FALSE</code> removes the box surrounding the graph.
<code>cex=1</code>	Character expansion. For example, when <code>cex=2</code> , characters are twice as big as normal.
<code>col=1</code>	Color used for drawing lines, points, etc.
<code>las=0</code>	The style of axis labels (0=always parallel to the axis—the default, 1=always horizontal, 2=always perpendicular to the axis, 3=always vertical)
<code>lty=1</code>	Line type (1=solid, 2=small breaks, etc.)
<code>lwd=1</code>	Line width (1=default, 2= twice as thick, etc.)
<code>main="title"</code>	Title for graph
<code>par()</code>	Used to set or query graphical parameters. See R or S-PLUS help files for more detail.
<code>pch=19</code>	Plotting symbol to use. For instance, 19 is a solid circle and 22 is a square
<code>pty="m"</code>	Type of plotting region: The default value for <code>pty</code> is <code>m</code> , which generates a maximal size plotting region. <code>pty="s"</code> generates a square plotting region.
<code>sub="subtitle"</code>	Subtitle for graph
<code>type="b"</code>	Both points and lines between points are used to represent data values.
<code>type="h"</code>	Height bars (vertical) represent data values.
<code>type="l"</code>	Lines are used to connect data values.
<code>type="p"</code>	Points are used to represent data values, the default argument.
<code>xlab="label"</code>	Label for x -axis
<code>xlim=c(xmin, xmax)</code>	Range for x -axis
<code>ylab="label"</code>	Label for y -axis
<code>ylim=c(ymin, ymax)</code>	Range for y -axis

Appendix B

Quadratic Forms and Random Vectors and Matrices

B.1 Quadratic Forms

DEFINITION B.1: Assume that the scalar W can be expressed as a function of the n variables Y_1, Y_2, \dots, Y_n . That is,

$$W = f(Y_1, Y_2, \dots, Y_n) = f(\mathbf{Y}) \text{ and } \frac{\delta W}{\delta \mathbf{Y}} = \begin{bmatrix} \frac{\delta W}{\delta Y_1} \\ \vdots \\ \frac{\delta W}{\delta Y_n} \end{bmatrix}.$$

DEFINITION B.2: Let \mathbf{A} be an $n \times n$ matrix and $\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$ be an $n \times 1$ column vector of real variables. Then $q = \mathbf{Y}'\mathbf{A}\mathbf{Y}$ is called a **quadratic form** in \mathbf{Y} , and \mathbf{A} is called the matrix of the quadratic form.

Rules for Differentiation

1. Let $W = \mathbf{A}'\mathbf{Y}$, where \mathbf{A} is a vector of scalars. Then $\frac{\delta W}{\delta \mathbf{Y}} = \mathbf{A}$.
2. Let $W = \mathbf{Y}'\mathbf{Y}$. Then, $\frac{\delta W}{\delta \mathbf{Y}} = 2\mathbf{Y}$.
3. Let $W = \mathbf{Y}'\mathbf{A}\mathbf{Y}$, where \mathbf{A} is an $n \times n$ matrix. Then $\frac{\delta W}{\delta \mathbf{Y}} = \mathbf{A}\mathbf{Y} + \mathbf{A}'\mathbf{Y}$.

Example B.1 Let $\mathbf{A} = \begin{bmatrix} 5 & 2 & 1 \\ 2 & 3 & -6 \\ 1 & -6 & 4 \end{bmatrix}$ and $\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix}$.

Then, $W = \mathbf{Y}'\mathbf{A}\mathbf{Y} = 5Y_1^2 + 3Y_2^2 + 4Y_3^2 + 4Y_1Y_2 + 2Y_1Y_3 - 12Y_2Y_3$, and the partial derivatives of W are

$$\begin{aligned} \frac{\delta W}{\delta Y_1} &= 10Y_1 + 4Y_2 + 2Y_3 \\ \frac{\delta W}{\delta Y_2} &= 6Y_2 + 4Y_1 - 12Y_3 \\ \frac{\delta W}{\delta Y_3} &= 8Y_3 + 2Y_1 - 12Y_2 \end{aligned}$$

or by Rule for Differentiation 3

$$\frac{\delta \mathbf{W}}{\delta \mathbf{Y}} = \mathbf{A}\mathbf{Y} + \mathbf{A}'\mathbf{Y} = \begin{bmatrix} 10Y_1 + 4Y_2 + 2Y_3 \\ 4Y_1 + 6Y_2 - 12Y_3 \\ 2Y_1 - 12Y_2 + 8Y_3 \end{bmatrix}$$

B.2 Random Vectors and Matrices

A random vector or a random matrix contains elements that are themselves random variables rather than real variables or scalar values.

DEFINITION B.3: Given a $p \times 1$ random vector $\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{bmatrix}$, the expected value of \mathbf{Y} ,

denoted by $E(\mathbf{Y})$, is defined as $E(\mathbf{Y}) = \begin{bmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_p) \end{bmatrix}$.

Basically, the expected value of a random vector is the vector of the expected values of the elements in the random vector. This concept extends to the expected value of a random matrix as well. That is, given a random $n \times p$ matrix \mathbf{Y} , $E(\mathbf{Y}) = [E(Y_{ij})]$ for all $i = 1, \dots, n$ and $j = 1, \dots, p$ pairs.

B.3 Variance of Random Vectors

Recall that the variance of a random variable Y defined in (3.7) on page 92 measures the variability of Y about its mean μ . Specifically,

$$\sigma_Y^2 = \text{Var}(Y) = E[(Y - E(Y))^2] = E[(Y - \mu)^2]$$

The notion of variability is slightly more challenging to extend to vectors and matrices. The difficulty arises because of the covariance between random variables. Recall that the covariance between random variables X and Y was defined as

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

in (5.14) on page 181. To compute both the variances and covariances of random variables in

the $p \times 1$ random vector $\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_p \end{bmatrix}$, we construct a $p \times p$ matrix where the diagonal entries

are the variances of Y_1 to Y_p , while the off diagonal entries are the covariances between Y_i and Y_j , where $i \neq j$.

DEFINITION B.4: The **variance-covariance** matrix of \mathbf{Y} , denoted $\sigma_{\mathbf{Y}}^2$, is defined as

$$\sigma_{\mathbf{Y}}^2 = E[(\mathbf{Y} - \mu_{\mathbf{Y}})(\mathbf{Y} - \mu_{\mathbf{Y}})'] \quad (\text{B.1})$$

The calculations of the expanded form of $\sigma_{\mathbf{Y}}^2$ are

$$\begin{aligned} \sigma_{\mathbf{Y}}^2 &= E[(\mathbf{Y} - \mu_{\mathbf{Y}})(\mathbf{Y} - \mu_{\mathbf{Y}})'] \\ &= E \left\{ \begin{bmatrix} Y_1 - \mu_1 \\ Y_2 - \mu_2 \\ \vdots \\ Y_p - \mu_p \end{bmatrix} [Y_1 - \mu_1, Y_2 - \mu_2, \dots, Y_p - \mu_p] \right\} \\ &= \begin{bmatrix} E[(Y_1 - \mu_1)^2] & E[(Y_1 - \mu_1)(Y_2 - \mu_2)] & \cdots & E[(Y_1 - \mu_1)(Y_p - \mu_p)] \\ E[(Y_2 - \mu_2)(Y_1 - \mu_1)] & E[(Y_2 - \mu_2)^2] & \cdots & E[(Y_2 - \mu_2)(Y_p - \mu_p)] \\ \vdots & \vdots & & \vdots \\ E[(Y_p - \mu_p)(Y_1 - \mu_1)] & E[(Y_p - \mu_p)(Y_2 - \mu_2)] & \cdots & E[(Y_p - \mu_p)^2] \end{bmatrix} \\ &= \begin{bmatrix} \sigma_{Y_1}^2 & \sigma_{Y_1 Y_2} & \cdots & \sigma_{Y_1 Y_p} \\ \sigma_{Y_2 Y_1} & \sigma_{Y_2}^2 & \cdots & \sigma_{Y_2 Y_p} \\ \vdots & \vdots & & \vdots \\ \sigma_{Y_p Y_1} & \sigma_{Y_p Y_2} & \cdots & \sigma_{Y_p}^2 \end{bmatrix} \end{aligned}$$

The following rules will help simplify complex expressions so that their variances can be determined more easily. It is frequently the case that a random vector, \mathbf{Z} , is obtained by premultiplying the random vector \mathbf{Y} by a constant matrix \mathbf{A} . That is, $\mathbf{Z} = \mathbf{A}\mathbf{Y}$.

1. $E[\mathbf{A}] = \mathbf{A}$
2. $E[\mathbf{Z}] = E[\mathbf{A}\mathbf{Y}] = \mathbf{A}E[\mathbf{Y}]$
3. $\sigma_{\mathbf{Z}}^2 = \sigma_{\mathbf{A}\mathbf{Y}}^2 = \mathbf{A}\sigma_{\mathbf{Y}}^2\mathbf{A}'$

where $\sigma_{\mathbf{Y}}^2$ is the variance-covariance matrix of \mathbf{Y} .

References

- Agresti, A. 1996. *An Introduction to Categorical Data Analysis*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. New York: John Wiley & Sons.
- . 2002. *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Hoboken, New Jersey: John Wiley & Sons.
- Agresti, A. and B. A. Coull. 1998. Approximate is better than *Exact* for interval estimation of binomial proportions. *The American Statistician* 52, no. 2:119–126.
- Arnholt, A. T. 2005. *BSDA: Basic Statistics and Data Analysis*. <http://www1.appstate.edu/~arnholta/software/>. R package version 0.1.
- . 2008a. *PASWR: Probability and Statistics with R*. <http://www1.appstate.edu/~arnholta/software/>. R package version 1.0.
- . 2008b. *PASWR: Probability and Statistics with R*. <http://www1.appstate.edu/~arnholta/software/>. S-PLUS package version 1.0.
- Artuch, R., C. Colomé, C. Sierra, N. Brandi, N. Lambruschini, J. Campistol, M. D. Ugarte, and M. Villaseca. 2004. Study of antioxidant status in phenylketonuric patients. *Clinical Biochemistry* 37:198–203.
- Belsley, D. A., E. Kuh, and R. E. Welsch. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley Series in Probability and Mathematical Statistics. New York: John Wiley & Sons.
- Blaker, H. 2000. Confidence Curves and Improved Exact Confidence Intervals for Discrete Distributions. *Canadian Journal of Statistics* 28:783–798.
- Blom, G. 1958. *Statistical Estimates and Transformed Beta-variables*. New York: John Wiley & Sons.
- Canty, A. 2007. *boot: Bootstrap R (S-Plus) Functions (Canty)*. S original by Angelo Canty <cantya@mcmaster.ca>. R port by Brian Ripley <ripley@stats.ox.ac.uk>, R package version 1.2-28.
- Cao, R., M. Francisco, S. Naya, M. Presedo, M. Vázquez, J. A. Vilar, and J. M. Vilar. 2001. *Introducción a la Estadística y sus Aplicaciones*. Madrid: Ediciones Pirámide.
- Casas, J. M., C. García, L. P. Rivera, and A. I. Zamora. 1998. *Problemas de Estadística*. Madrid: Ediciones Pirámide.
- Casella, G. and R. L. Berger. 1990. *Statistical Inference*. The Wadsworth & Brooks/Cole Statistics/Probability Series. Pacific Grove, California: Wadsworth & Brooks/Cole Advanced Books & Software.
- Chatterjee, S. and A. S. Hadi. 1988. *Sensitivity Analysis in Linear Regression*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. New York: John Wiley & Sons.

- Chatterjee, S. and B. Price. 1991. *Regression Diagnostics*. New York: John Wiley & Sons.
- Christensen, R. 1996. *Analysis of Variance, Design and Regression*. New York: Chapman and Hall.
- Chu, S. 2003. Using Soccer Goals to Motivate the Poisson Process. *INFORMS Transaction on Education* 3, no. 2:62–68.
- Cleveland, W. S. 1993. *Visualizing Data*. Summit, New Jersey: Hobart Press.
- . 1994. *The Elements of Graphing Data*. Summit, New Jersey: Hobart Press.
- Cochran, W. G. 1977. *Sampling Techniques*. Wiley Series in Probability and Mathematical Statistics. New York: John Wiley & Sons, third ed.
- Conover, W. J. 1999. *Practical Nonparametric Statistics*. Monographs on Statistics and Applied Probability. New York: John Wiley & Sons, third ed.
- Cook, R. D. and S. Weisberg. 1982. *Residuals and Influence in Regression*. Monographs on Statistics and Applied Probability. London: Chapman & Hall.
- Cuesta, M. J., M. D. Ugarte, T. Goicoa, S. Eraso, and V. Peralta. 2007. A Taxometric Analysis of Schizophrenia Symptoms. *Psychiatry Research* 150:245–253.
- Dahl, D. B. 2006. *xtable: Export tables to L^AT_EX or HTML*. R package version 1.4-2.
- Dalgaard, P. 2002. *Introductory Statistics with R*. Statistics and Computing. New York: Springer-Verlag.
- Dallal, G. E. and L. Wilkinson. 1986. An analytic approximation to the distribution of Lilliefors' test for normality. *The American Statistician* 40:294–296.
- Davis, J. 1986. *Statistics and Data Analysis in Geology*. New York: John Wiley & Sons.
- Davison, A. C. and D. V. Hinkley. 1997. *Bootstrap Methods and Their Application*, vol. 1 of Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press. With 1 IBM-PC floppy disk (3.5 inch; HD).
- Devore, J. L. 2000. *Probability and Statistics for Engineering and the Sciences*. Monographs on Statistics and Applied Probability. Australia: Duxbury Thomson Learning, fifth ed.
- Draper, N. R. and H. Smith. 1998. *Applied Regression Analysis*. Wiley Series in Probability and Statistics: Texts and References Section. New York: John Wiley & Sons, third ed. With 1 IBM-PC floppy disk (3.5 inch; DD).
- Efron, B. 1979. Bootstrap methods: another look at the jackknife. *Annals of Statistics* 7:1–26.
- Efron, B. and R. J. Tibshirani. 1993. *An Introduction to the Bootstrap*, vol. 57 of Monographs on Statistics and Applied Probability. New York: Chapman and Hall.
- Faraway, J. J. 2005. *Linear models with R*. Chapman & Hall/CRC Texts in Statistical Science Series. Boca Raton, Florida: Chapman & Hall/CRC.
- . 2006. *faraway: Functions and data sets for books by Julian Faraway*. <http://www.maths.bath.ac.uk/~jjf23/>. R package version 1.0.0.
- Fdez. Militino, A., S. Gómez, and G. Aldaz. 1994. *Problemas Resueltos y Aplicaciones de Estadística*. Pamplona: UNED.
- Fisher, R. A. 1936. The Use of Multiple Measurements in Taxonomic Problems. *Annals of*

- Eugenics* 7, no. II:179–188.
- . 1971. *The Design of Experiments*. New York: Macmillan Publishing Co., ninth ed.
- Fleiss, J. L. 1981. *Statistical Methods for Rates and Proportions*. Wiley Series in Probability and Mathematical Statistics. New York: John Wiley & Sons, second ed.
- Fox, J. 1997. *Applied Regression Analysis, Linear Models, and Related Methods*. Thousand Oaks, California: SAGE Publications, first ed.
- . 2002. *An R and S-Plus Companion to Applied Regression*. Thousand Oaks, California: SAGE Publications, first ed.
- . 2006. *car: Companion to Applied Regression*. <http://www.r-project.org>, <http://socserv.socsci.mcmaster.ca/jfox/>. R package version 1.2-1.
- Gibbons, J. D. 1997. *Nonparametric Methods for Quantitative Analysis*, vol. 2 of American Series in Mathematical and Management Sciences. Columbus, Ohio: American Sciences Press, third ed.
- Gibbons, J. D. and S. Chakraborti. 1992. *Nonparametric Statistical Inference*, vol. 131 of Statistics: Textbooks and Monographs. New York: Marcel Dekker, third ed.
- . 2003. *Nonparametric Statistical Inference*, vol. 168 of Statistics: Textbooks and Monographs. New York: Marcel Dekker, fourth ed.
- Gilmour, S. G. 1996. The Interpretation of Mallows's C_P -Statistic. *The Statistician* 45, no. 1:49–56.
- Graves, S. and H.-P. Piepho. 2006. *multcompView: Visualizations of Paired Comparisons*. With help from Sundar Dorai-Raj. R package version 0.1-0.
- Gray, J. B. and H. Woodall. 1994. The Maximum Size of Standardized and Internally Studentized Residuals in Regression Analysis. *The American Statistician* 48, no. 2:111–113.
- Graybill, F. A. 1976. *Theory and Application of the Linear Model*. Belmont, California: Wadsworth Publishing Company.
- Gross, J. 2003. *nortest: Tests for Normality*. R package version 1.0.
- Hand, D. J., F. Daly, A. D. Lunn, K. J. McConway, and E. Ostrowski. 1994. *A Handbook of Small Data Sets*. London: Chapman & Hall.
- Harrell, F. E. 2006. *Hmisc: Harrell Miscellaneous*. <http://biostat.mc.vanderbilt.edu/s/Hmisc>. R package version 3.1-2.
- Hennekens, C. 1988. Preliminary Report: Findings from the Aspirin Component of the Ongoing Physician's Health Study. *New England Journal of Medicine* 318:262–264.
- Hicks, C. R. 1956. Fundamentals of analysis of variance, part ii—the components of variance model and the mixed model. *Industrial Quality Control* 13:5–8.
- Hines, W. G. S. 1996. Pragmatics of Pooling in ANOVA Tables. *The American Statistician* 50, no. 2:127–139.
- Hines, W. W. and D. C. Montgomery. 1990. *Probability and Statistics in Engineering and Management Science*. New York: John Wiley & Sons, third ed.

- Hocking, R. R. 1996. *Methods and Applications of Linear Models*. New York: John Wiley & Sons.
- Hollander, M. and D. A. Wolfe. 1999. *Nonparametric Statistical Methods*. Wiley Series in Probability and Statistics: Texts and References Section. New York: John Wiley & Sons, second ed.
- Hothorn, T. and K. Hornik. 2006. *exactRankTests: Exact Distributions for Rank and Permutation Tests*. R package version 0.8-15.
- Hothorn, T., K. Hornik, M. A. van de Wiel, and A. Zeileis. 2006. A Lego system for conditional inference. *The American Statistician* 60, no. 3:257–263.
- Hothorn, T., F. Bretz, and P. Westfall. 2007. *multcomp: Simultaneous Inference for General Linear Hypotheses*. With contributions by Richard M. Heiberger. R package version 0.992-4.
- Insightful Corporation. 2005a. *S-PLUS 7 Application Developer's Guide*. Seattle, WA. <http://www.insightful.com>.
- . 2005b. *S-PLUS 7 for Windows User's Guide*. Seattle, WA. <http://www.insightful.com>.
- . 2005c. *S-PLUS 7 Guide to Statistics, Volume 1*. Seattle, WA. <http://www.insightful.com>.
- . 2005d. *S-PLUS 7 Guide to Statistics, Volume 2*. Seattle, WA. <http://www.insightful.com>.
- . 2005e. *S-PLUS 7 Programmer's Guide*. Seattle, WA. <http://www.insightful.com>.
- . 2007. *S-PLUS 8 Guide to Packages*. Seattle, WA. <http://www.insightful.com>.
- Johnson, N. L., S. Kotz, and N. Balakrishnan. 1995. *Continuous Univariate Distributions. Vol. 2*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. New York: John Wiley & Sons, second ed.
- Kalbfleisch, J. G. 1985a. *Probability and Statistical Inference. Vol. 1*. Springer Texts in Statistics. New York: Springer-Verlag, second ed.
- . 1985b. *Probability and Statistical Inference. Vol. 2*. Springer Texts in Statistics. New York: Springer-Verlag, second ed.
- Kitchens, L. J. 2003. *Basic Statistics and Data Analysis*. Pacific Grove, California: Brooks/Cole, a division of Thomson Learning.
- Kleinbaum, D. and L. Kupper. 1998. *Applied Regression Analysis and Other Multivariable Methods*. London: Duxbury Press, third ed.
- Kopka, H. and P. W. Daly. 1995. *A Guide to L^AT_EX 2_ε*. New York: Addison-Wesley, second ed.
- Krause, A. and M. Olson. 1997. *The Basics of S and S-PLUS*. New York: Springer-Verlag.
- Kutner, M. H., C. J. Nachtsheim, J. Neter, and W. Li. 2004. *Applied Linear Statistical Models*. Boston: McGraw-Hill/Irwin, fifth ed.
- Lapin, L. L. 1990. *Probability and Statistics for Modern Engineering*. Boston: PWS-KENT Publishing Company, second ed.

- Lawless, J. 1982. *Statistical Models and Methods for Lifetime Data*. New York: John Wiley & Sons.
- Levy, P. S. and S. Lemeshow. 1999. *Sampling of Populations*. Boston: John Wiley & Sons, third ed.
- Lilliefors, H. W. 1967. On the Kolmogorov-Smirnov Tests for Normality with Mean and Variance Unknown. *Journal of the American Statistical Association* 62:399–402.
- López, J. 1994. *Problemas de Inferencia Estadística, (Muestreo y Control de Calidad)*. Albacete: Tébar Flores, third ed.
- Lumley, T. 2004. *Leaps: Regression Subset Selection*. R package version 2.7 using Fortran code by Alan Miller.
- Lunneborg, C. E. 2000. *Data Analysis by Resampling: Concepts and Applications*. Pacific Grove, California: Duxbury, first ed.
- Maindonald, J. and J. Braun. 2003. *Data Analysis and Graphics Using R—An Example-Based Approach*, vol. 10 of Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.
- Mazess, R. B., W. W. Peppler, and M. Gibbons. 1984. Total Body Composition by Dual-Photon (^{153}Gd) Absorptiometry. *American Journal of Clinical Nutrition* 40, no. 4:834–839.
- Militino, A. F., M. D. Ugarte, T. Goicoa, and M. González-Audicana. 2006. Using Small Area Models to Estimate the Total Area Occupied by Olive Trees. *Journal of Agricultural, Biological and Environmental Statistics* 11:450–461.
- Montgomery, D. C. 1991. *Design and Analysis of Experiments*. New York: John Wiley & Sons, third ed.
- Muro, J., I. Irigoyen, A. Militino, and C. Lamsfus. 2001. Defoliation Effects on Sunflower Yield Reduction. *Agronomy Journal* 93:634–637.
- Murrell, P. 2006. *R Graphics*. Computer Science and Data Analysis Series. Boca Raton, Florida: Chapman & Hall/CRC.
- Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. 1996. *Applied Linear Statistical Models*. Boston: McGraw-Hill, fourth ed.
- Oehlert, G. W. 2000. *A First Course in Design and Analysis of Experiments*. New York: W. H. Freeman and Company.
- Ott, L. and W. Mendenhall. 1985. *Understanding Statistics*. Boston: Duxbury Press.
- Peña, D. 2001. *Fundamentos de Estadística*. Madrid: Alianza Editorial.
- . 2002. *Regresión y Diseño de Experimentos*. Madrid: Alianza Editorial.
- Petrucelli, J. D., B. Nandram, and M. Chen. 1999. *Applied Statistics for Engineers and Scientists*. Upper Saddle River, New Jersey: Prentice Hall.
- Pinheiro, J. C. and D. M. Bates. 2000. *Mixed-Effects Models in S and S-PLUS*. New York: Springer-Verlag.
- Pratt, J. W. and J. D. Gibbons. 1981. *Concepts of Nonparametric Theory*. New York: Springer-Verlag.

- R Development Core Team. 2006. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org>.
- . 2007a. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org>.
- . 2007b. *R: R Data Import/Export*. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org>.
- . 2007c. *R: R Installation and Administration*. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org>.
- . 2007d. *R: Writing R Extensions*. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org>.
- Rao, C. R. 1973. *Linear Statistical Inference and Its Applications*. New York: John Wiley & Sons.
- Ross, S. 1994. *A First Course in Probability*. Upper Saddle River, New Jersey: Prentice Hall, fifth ed.
- Rousseeuw, P. J. and B. C. V. Zomeren. 1990. Unmasking Multivariate Outliers and Leverage Points. *Journal of the American Statistical Association* 85, no. 411:633–639.
- Ruiz-Maya, L. 1986. *Métodos Estadísticos de Investigación*. Madrid: Editorial INE, second ed.
- Ruiz-Maya, L. and F. J. Martín Pliego. 2001. *Estadística. II Inferencia*. Madrid: Editorial AC, second ed.
- Ryan, T. P. 1997. *Modern Regression Methods*. New York: John Wiley & Sons.
- Sarkar, D. 2007. *lattice: Lattice Graphics*. R package version 0.15-11.
- Scheaffer, R. L., W. Mendenhall, and L. Ott. 1986. *Elementos de Muestreo*. Mexico: Grupo Editorial Iberoamérica.
- Scott, D. W. 1992. *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: John Wiley & Sons.
- Seber, G. A. F. 1977. *Linear Regression Analysis*. New York: John Wiley & Sons.
- Selvin, S. 1998. *Modern Applied Biostatistical Methods Using S-PLUS*. New York: Oxford University Press.
- Shapiro, S. S. and M. B. Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52, no. 3 and 4:591–611.
- Sheskin, D. J. 1997. *Parametric and Nonparametric Statistical Procedures*. New York: CRC Press.
- Singh, R. and S. N. 1996. *Elements of Survey Sampling*. Dordrecht: Kluwer Academic Publishers.
- Sokal, R. R. and F. J. Rohlf. 1994. *Biometry*. New York: W. H. Freeman, third ed.
- Sternberg, D. E., D. P. Van Kammen, and W. E. Bunney. 1982. Schizophrenia: Dopamine b-Hydroxylase Activity and Treatment Response. *Science* 216:1423–1425.
- Tibshirani, R. 2006. *bootstrap: Functions for the Book "An Introduction to the Bootstrap"*.

- S original by Rob Tibshirani. R port by Friedrich Leisch. R package version 1.0-20.
- Ugarte, M. D. and A. F. Militino. 2002. *Estadística Aplicada con S-PLUS*. Spain: Universidad Pública de Navarra, second ed.
- Venables, W. N. and B. D. Ripley. 1999. *Modern Applied Statistics with S-PLUS*. New York: Springer-Verlag, third ed.
- . 2000. *S Programming*. New York: Springer-Verlag.
- . 2002. *Modern Applied Statistics with S*. New York: Springer-Verlag, fourth ed. <http://www.stats.ox.ac.uk/pub/MASS4>.
- Verzani, J. 2005. *Using R for Introductory Statistics*. Boca Raton, Florida: Chapman & Hall/CRC.
- Warnes, G. R. 2006. *gregmisc: Greg's Miscellaneous Functions*. R package version 2.1.1.
- Weindling, A. M., F. N. Bamford, and R. A. Whittall. 1986. Health of Juvenile Delinquents. *British Medical Journal* 292:447.
- Welch, B. L. 1951. On the comparison of several mean values: an alternative approach. *Biometrika* 38:330–336.
- Yandell, B. S. 1997. *Practical Data Analysis for Designed Experiments*. New York: Chapman & Hall.
- Zaman, A. 2000. The Inconsistency of the Breusch-Pagan Test. *Journal of Economic and Social Research* 2:1–11.
- Zeileis, A. and T. Hothorn. 2002. Diagnostic checking in regression relationships. *R News* 2, no. 3:7–10. <http://CRAN.R-project.org/doc/Rnews/>.

PROBABILITY *and* STATISTICS WITH R

Practical and Visually Appealing with Clear Examples and Fully Detailed Proofs

Probability and Statistics with R shows how to solve various statistical problems using both parametric and nonparametric techniques via the open source software R. It provides numerous real-world examples, carefully explained proofs, end-of-chapter problems, and illuminating graphs to facilitate hands-on comprehension.

Delves into Many Probability and Statistical Topics

Integrating theory with practice, the text briefly introduces the syntax, structures, and functions of the S language, before covering important graphically and numerically descriptive methods. The next several chapters elucidate probability and random variables topics, including univariate and multivariate distributions. After exploring sampling distributions, the authors discuss point estimation, confidence intervals, hypothesis testing, and a wide range of nonparametric methods. With a focus on experimental design, the book also presents fixed- and random-effects models as well as randomized block and two-factor factorial designs. The final chapter describes simple and multiple regression analyses.

Cohesively Incorporates Statistical Theory with R Implementation

This comprehensive book presents extensive treatments of data analysis using parametric and nonparametric techniques. It effectively links statistical concepts with R procedures, enabling the application of the language to the vast world of statistics.

Features

- Provides real-world examples of how R can be used to solve problems in probability and statistics, along with an overview on how to use these computer languages
- Explains the mathematics behind computational implementations
- Covers both traditional methods and nonparametric techniques, including goodness-of-fit tests, categorical data analysis, nonparametric bootstrapping, and permutation tests
- Uses regression analysis procedures to solve three interesting case studies based on real data
- Presents thoroughly worked-out derivations, detailed graphs, and abundant problems
- Offers data sets, functions, and other ancillary material on a supporting website

C8911

ISBN: 978-1-58488-891-8

90000



9 781584 888918

www.crcpress.com



CRC Press

Taylor & Francis Group
an Informa business

www.crcpress.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
270 Madison Avenue
New York, NY 10016
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK