

The Statistical Sleuth

A Course in Methods of Data Analysis

THIRD EDITION

Ramsey/Schafer





The Statistical Sleuth

A Course in Methods of Data Analysis

THIRD EDITION

Fred L. Ramsey
Oregon State University

Daniel W. Schafer
Oregon State University



Australia • Canada • Mexico • Singapore • Spain • United Kingdom • United States

This is an electronic version of the print textbook. Due to electronic rights restrictions, some third party content may be suppressed. Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. The publisher reserves the right to remove content from this title at any time if subsequent rights restrictions require it. For valuable information on pricing, previous editions, changes to current editions, and alternate formats, please visit www.cengage.com/highered to search by ISBN#, author, title, or keyword for materials in your areas of interest.

**The Statistical Sleuth: A Course in Methods of
Data Analysis, Third Edition**

Fred L. Ramsey, Daniel W. Schafer

Publisher: Richard Stratton

Senior Sponsoring Editor: Molly Taylor

Assistant Editor: Shaylin Walsh Hogan

Editorial Assistant: Alexander Gontar

Associate Media Editor: Andrew Coppola

Senior Marketing Manager: Barb Bartoszek

Marketing Assistant: Lindsay M Lettre

Marketing Communications Manager:

Mary Anne Payumo

Manufacturing Planner: Doug Bertke

Rights Acquisitions Specialist:

Shalice Shah-Caldwell

Design Direction, Production Management, and
Composition: PreMediaGlobal

Cover Image: © Pali Rao/iStockphoto.com

© 2013, 2002 Brooks/Cole, Cengage Learning

ALL RIGHTS RESERVED. No part of this work covered by the copyright herein may be reproduced, transmitted, stored, or used in any form or by any means graphic, electronic, or mechanical, including but not limited to photocopying, recording, scanning, digitizing, taping, Web distribution, information networks, or information storage and retrieval systems, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without the prior written permission of the publisher.

For product information and technology assistance, contact us at
Cengage Learning Customer & Sales Support, 1-800-354-9706

For permission to use material from this text or product,
submit all requests online at www.cengage.com/permissions.

Further permissions questions can be emailed to
permissionrequest@cengage.com

Library of Congress Control Number: 2012931469

Student Edition:

ISBN-13: 978-1-133-49067-8

ISBN-10: 1-133-49067-0

Brooks/Cole

20 Channel Center Street
Boston, MA 02210
USA

Cengage Learning is a leading provider of customized learning solutions with office locations around the globe, including Singapore, the United Kingdom, Australia, Mexico, Brazil and Japan. Locate your local office at
international.cengage.com/region

Cengage Learning products are represented in Canada by Nelson Education, Ltd.

For your course and learning solutions, visit www.cengage.com

Purchase any of our products at your local college store or at our preferred online store www.cengagebrain.com

Instructors: Please visit login.cengage.com and log in to access instructor-specific resources.

All Displays in this edition are owned by Cengage Learning. © 2013 Cengage Learning

Printed in the United States of America
1 2 3 4 5 6 7 16 15 14 13 12

Dedications

To influential teachers Don Truax and Bob Buehler—F.L.R

To Jeannie, Banner, and Casey—D.S.



Contents

Preface

xx

CHAPTER 1 Drawing Statistical Conclusions

1

1.1	Case Studies 2
1.1.1	Motivation and Creativity—A Randomized Experiment 2
1.1.2	Sex Discrimination in Employment—An Observational Study 4
1.2	Statistical Inference and Study Design 5
1.2.1	Causal Inference 5
1.2.2	Inference to Populations 7
1.2.3	Statistical Inference and Chance Mechanisms 8
1.3	Measuring Uncertainty in Randomized Experiments 10
1.3.1	A Probability Model for Randomized Experiments 10
1.3.2	A Test for Treatment Effect in the Creativity Study 11
1.4	Measuring Uncertainty in Observational Studies 14
1.4.1	A Probability Model for Random Sampling 14
1.4.2	Testing for a Difference in the Sex Discrimination Study 15
1.5	Related Issues 16
1.5.1	Graphical Methods 16
1.5.2	Standard Statistical Terminology 19
1.5.3	Randomization of Experimental Units to Treatments 21
1.5.4	Selecting a Simple Random Sample from a Population 21
1.5.5	On Being Representative 22
1.6	Summary 22
1.7	Exercises 22
	Conceptual Exercises 22
	Computational Exercises 24
	Data Problems 25
	Answers to Conceptual Exercises 26

CHAPTER 2 Inference Using *t*-Distributions

28

2.1	Case Studies 29
2.1.1	Evidence Supporting Darwin’s Theory of Natural Selection—An Observational Study 29
2.1.2	Anatomical Abnormalities Associated with Schizophrenia—An Observational Study 31
2.2	One-Sample <i>t</i> -Tools and the Paired <i>t</i> -Test 32
2.2.1	The Sampling Distribution of a Sample Average 32

2.2.2	The Standard Error of an Average in Random Sampling	34
2.2.3	The <i>t</i> -Ratio Based on a Sample Average	35
2.2.4	Unraveling the <i>t</i> -Ratio	36
2.3	A <i>t</i> -Ratio for Two-Sample Inference	38
2.3.1	Sampling Distribution of the Difference Between Two Independent Sample Averages	39
2.3.2	Standard Error for the Difference of Two Averages	40
2.3.3	Confidence Interval for the Difference Between Population Means	42
2.3.4	Testing a Hypothesis About the Difference Between Means	43
2.3.5	The Mechanics of <i>p</i> -Value Computation	45
2.4	Inferences in a Two-Treatment Randomized Experiment	45
2.4.1	Approximate Uncertainty Measures for Randomized Experiments	46
2.5	Related Issues	47
2.5.1	Interpretation of <i>p</i> -Values	47
2.5.2	An Example of Confidence Intervals	49
2.5.3	The Rejection Region Approach to Hypothesis Testing	50
2.6	Summary	51
2.7	Exercises	51
	Conceptual Exercises	51
	Computational Exercises	52
	Data Problems	54
	Answers to Conceptual Exercises	56

CHAPTER 3 A Closer Look at Assumptions 58

3.1	Case Studies	59
3.1.1	Cloud Seeding to Increase Rainfall—A Randomized Experiment	59
3.1.2	Effects of Agent Orange on Troops in Vietnam—An Observational Study	60
3.2	Robustness of the Two-Sample <i>t</i> -Tools	62
3.2.1	The Meaning of Robustness	62
3.2.2	Robustness Against Departures from Normality	62
3.2.3	Robustness Against Differing Standard Deviations	64
3.2.4	Robustness Against Departures from Independence	65
3.3	Resistance of the Two-Sample <i>t</i> -Tools	66
3.3.1	Outliers and Resistance	66
3.3.2	Resistance of <i>t</i> -Tools	67
3.4	Practical Strategies for the Two-Sample Problem	67
3.5	Transformations of the Data	69
3.5.1	The Logarithmic Transformation	69
3.5.2	Interpretation After a Log Transformation	71
3.5.3	Other Transformations for Positive Measurements	74
3.6	Related Issues	75
3.6.1	Prefer Graphical Methods over Formal Tests for Model Adequacy	75
3.6.2	Robustness and Transformation for Paired <i>t</i> -Tools	76
3.6.3	Example—Schizophrenia	76
3.7	Summary	76
3.8	Exercises	77
	Conceptual Exercises	77
	Computational Exercises	79
	Data Problems	82
	Answers to Conceptual Exercises	84

CHAPTER 4 Alternatives to the <i>t</i>-Tools	85
4.1 Case Studies	86
4.1.1 Space Shuttle O-Ring Failures—An Observational Study	86
4.1.2 Cognitive Load Theory in Teaching—A Randomized Experiment	87
4.2 The Rank-Sum Test	88
4.2.1 The Rank Transformation	89
4.2.2 The Rank-Sum Statistic	90
4.2.3 Finding a <i>p</i> -Value by Normal Approximation	90
4.2.4 A Confidence Interval Based on the Rank-Sum Test	94
4.3 Other Alternatives for Two Independent Samples	95
4.3.1 Permutation Tests	95
4.3.2 The Welch <i>t</i> -Test for Comparing Two Normal Populations with Unequal Spreads	97
4.4 Alternatives for Paired Data	99
4.4.1 The Sign Test	99
4.4.2 The Wilcoxon Signed-Rank Test	100
4.5 Related Issues	102
4.5.1 Practical and Statistical Significance	102
4.5.2 The Presentation of Statistical Findings	103
4.5.3 Levene's (Median) Test for Equality of Two Variances	103
4.5.4 Survey Sampling	104
4.6 Summary	105
4.7 Exercises	106
Conceptual Exercises	106
Computational Exercises	106
Data Problems	109
Answers to Conceptual Exercises	111
CHAPTER 5 Comparisons Among Several Samples	113
5.1 Case Studies	114
5.1.1 Diet Restriction and Longevity—A Randomized Experiment	114
5.1.2 The Spock Conspiracy Trial—An Observational Study	117
5.2 Comparing Any Two of the Several Means	119
5.2.1 An Ideal Model for Several-Sample Comparisons	119
5.2.2 The Pooled Estimate of the Standard Deviation	120
5.2.3 <i>t</i> -Tests and Confidence Intervals for Differences of Means	120
5.3 The One-Way Analysis of Variance <i>F</i>-Test	121
5.3.1 The Extra-Sum-of-Squares Principle	122
5.3.2 The Analysis of Variance Table for One-Way Classification	126
5.4 More Applications of the Extra-Sum-of-Squares <i>F</i>-Test	127
5.4.1 Example: Testing Equality in a Subset of Groups	127
5.4.2 Summary of ANOVA Tests Involving More Than Two Models	129
5.5 Robustness and Model Checking	130
5.5.1 Robustness to Assumptions	130
5.5.2 Diagnostics Using Residuals	131
5.6 Related Issues	133
5.6.1 Further Illustration of the Different Sources of Variability	133
5.6.2 Kruskal–Wallis Nonparametric Analysis of Variance	136
5.6.3 Random Effects	137
5.6.4 Separate Confidence Intervals and Significant Differences	139

5.7	Summary	140
5.8	Exercises	141
	Conceptual Exercises	141
	Computational Exercises	142
	Data Problems	146
	Answers to Conceptual Exercises	147

CHAPTER 6 Linear Combinations and Multiple Comparisons of Means 149

6.1	Case Studies	150
6.1.1	Discrimination Against the Handicapped—A Randomized Experiment	150
6.1.2	Pre-Existing Preferences of Fish—A Randomized Experiment	151
6.2	Inferences About Linear Combinations of Group Means	152
6.2.1	Linear Combinations of Group Means	152
6.2.2	Inferences About Linear Combinations of Group Means	154
6.2.3	Specific Linear Combinations	155
6.3	Simultaneous Inferences	159
6.4	Some Multiple Comparison Procedures	161
6.4.1	Tukey–Kramer Procedure and the Studentized Range Distributions	161
6.4.2	Dunnett’s Procedure	162
6.4.3	Scheffé’s Procedure	162
6.4.4	Other Multiple Comparisons Procedures	162
6.4.5	Multiple Comparisons in the Handicap Study	164
6.4.6	Choosing a Multiple Comparisons Procedure	165
6.5	Related Issues	165
6.5.1	Reasoning Fallacies Associated with Statistical Hypothesis Testing and <i>p</i> -Values	165
6.5.2	Example of a Hypothesis Based on How the Data Turned Out	165
6.5.3	Is Choosing a Transformation a Form of Data Snooping?	169
6.6	Summary	169
6.7	Exercises	170
	Conceptual Exercises	170
	Computational Exercises	171
	Data Problems	173
	Answers to Conceptual Exercises	175

CHAPTER 7 Simple Linear Regression: A Model for the Mean 176

7.1	Case Studies	177
7.1.1	The Big Bang—An Observational Study	177
7.1.2	Meat Processing and pH—A Randomized Experiment	179
7.2	The Simple Linear Regression Model	180
7.2.1	Regression Terminology	180
7.2.2	Interpolation and Extrapolation	181
7.3	Least Squares Regression Estimation	183
7.3.1	Fitted Values and Residuals	183
7.3.2	Least Squares Estimators	184
7.3.3	Sampling Distributions of the Least Squares Estimators	184
7.3.4	Estimation of σ from Residuals	184
7.3.5	Standard Errors	186

7.4	Inferential Tools	188
7.4.1	Tests and Confidence Intervals for Slope and Intercept	188
7.4.2	Describing the Distribution of the Response Variable at Some Value of the Explanatory Variable	189
7.4.3	Prediction of a Future Response	191
7.4.4	Calibration: Estimating the X That Results in $Y = Y_0$	192
7.5	Related Issues	194
7.5.1	Historical Notes About Regression	194
7.5.2	Differing Terminology	196
7.5.3	Causation	196
7.5.4	Correlation	196
7.5.5	Planning an Experiment: Replication	197
7.6	Summary	197
7.7	Exercises	198
	Conceptual Exercises	198
	Computational Exercises	199
	Data Problems	202
	Answers to Conceptual Exercises	206

CHAPTER 8 A Closer Look at Assumptions for Simple Linear Regression 207

8.1	Case Studies	208
8.1.1	Island Area and Number of Species—An Observational Study	208
8.1.2	Breakdown Times for Insulating Fluid Under Different Voltages—A Controlled Experiment	209
8.2	Robustness of Least Squares Inferences	211
8.3	Graphical Tools for Model Assessment	213
8.3.1	Scatterplot of the Response Variable Versus the Explanatory Variable	213
8.3.2	Scatterplots of Residuals Versus Fitted Values	215
8.4	Interpretation After Log Transformations	216
8.5	Assessment of Fit Using the Analysis of Variance	218
8.5.1	Three Models for the Population Means	218
8.5.2	The Analysis of Variance Table Associated with Simple Regression	218
8.5.3	The Lack-of-Fit F -Test	220
8.5.4	A Composite Analysis of Variance Table	220
8.6	Related Issues	222
8.6.1	R -Squared: The Proportion of Variation Explained	222
8.6.2	Simple Linear Regression or One-Way Analysis of Variance?	223
8.6.3	Other Residual Plots for Special Situations	224
8.6.4	Planning an Experiment: Balance	225
8.7	Summary	226
8.8	Exercises	227
	Conceptual Exercises	227
	Computational Exercises	229
	Data Problems	231
	Answers to Conceptual Exercises	235

CHAPTER 9 Multiple Regression 237

9.1	Case Studies	238
9.1.1	Effects of Light on Meadowfoam Flowering—A Randomized Experiment	238

9.1.2	Why Do Some Mammals Have Large Brains for Their Size?—An Observational Study	239
9.2	Regression Coefficients	242
9.2.1	The Multiple Linear Regression Model	242
9.2.2	Interpretation of Regression Coefficients	243
9.3	Specially Constructed Explanatory Variables	246
9.3.1	A Squared Term for Curvature	246
9.3.2	An Indicator Variable to Distinguish Between Two Groups	247
9.3.3	Sets of Indicator Variables for Categorical Explanatory Variables with More Than Two Categories	249
9.3.4	A Product Term for Interaction	250
9.3.5	A Shorthand Notation for Model Description	252
9.4	A Strategy for Data Analysis	253
9.5	Graphical Methods for Data Exploration and Presentation	254
9.5.1	A Matrix of Pairwise Scatterplots	254
9.5.2	Coded Scatterplots	257
9.5.3	Jittered Scatterplots	257
9.5.4	Trellis Graphs	257
9.6	Related Issues	258
9.6.1	Computer Output	258
9.6.2	Factorial Treatment Arrangement	259
9.7	Summary	260
9.8	Exercises	261
	Conceptual Exercises	261
	Computational Exercises	263
	Data Problems	267
	Answers to Conceptual Exercises	269

CHAPTER 10 Inferential Tools for Multiple Regression 271

10.1	Case Studies	272
10.1.1	Galileo’s Data on the Motion of Falling Bodies—A Controlled Experiment	272
10.1.2	The Energy Costs of Echolocation by Bats—An Observational Study	273
10.2	Inferences About Regression Coefficients	275
10.2.1	Least Squares Estimates and Standard Errors	276
10.2.2	Tests and Confidence Intervals for Single Coefficients	277
10.2.3	Tests and Confidence Intervals for Linear Combinations of Coefficients	280
10.2.4	Prediction	283
10.3	Extra-Sums-of-Squares <i>F</i> -Tests	284
10.3.1	Comparing Sizes of Residuals in Hierarchical Models	285
10.3.2	<i>F</i> -Test for the Joint Significance of Several Terms	285
10.3.3	The Analysis of Variance Table	287
10.4	Related Issues	289
10.4.1	Further Notes on the <i>R</i> -Squared Statistic	289
10.4.2	Improving Galileo’s Design with Replication	292
10.4.3	Variance Formulas for Linear Combinations of Regression Coefficients	293
10.4.4	Further Notes About Polynomial Regression	295

10.4.5	Finding Where the Mean Response Is at Its Maximum in Quadratic Regression	295
10.4.6	The Principle of Occam's Razor	296
10.4.7	Informal Tests in Model Fitting	296
10.5	Summary	296
10.6	Exercises	297
	Conceptual Exercises	297
	Computational Exercises	299
	Data Problems	304
	Answers to Conceptual Exercises	309
CHAPTER 11	Model Checking and Refinement	310
11.1	Case Studies	311
11.1.1	Alcohol Metabolism in Men and Women—An Observational Study	311
11.1.2	The Blood–Brain Barrier—A Controlled Experiment	313
11.2	Residual Plots	316
11.3	A Strategy for Dealing with Influential Observations	319
11.4	Case-Influence Statistics	322
11.4.1	Leverages for Flagging Cases with Unusual Explanatory Variable Values	322
11.4.2	Studentized Residuals for Flagging Outliers	325
11.4.3	Cook's Distances for Flagging Influential Cases	325
11.4.4	A Strategy for Using Case Influence Statistics	326
11.5	Refining the Model	328
11.5.1	Testing Terms	328
11.5.2	Partial Residual Plots	329
11.6	Related Issues	334
11.6.1	Weighted Regression for Certain Types of Nonconstant Variance	334
11.6.2	The Delta Method	335
11.6.3	Measurement Errors in Explanatory Variables	335
11.7	Summary	337
11.8	Exercises	338
	Conceptual Exercises	338
	Computational Exercises	339
	Data Problems	342
	Answers to Conceptual Exercises	343
CHAPTER 12	Strategies for Variable Selection	345
12.1	Case Studies	346
12.1.1	State Average SAT Scores—An Observational Study	346
12.1.2	Sex Discrimination in Employment—An Observational Study	348
12.2	Specific Issues Relating to Many Explanatory Variables	352
12.2.1	Objectives	352
12.2.2	Loss of Precision	354
12.2.3	A Strategy for Dealing with Many Explanatory Variables	355
12.3	Sequential Variable-Selection Techniques	357
12.3.1	Forward Selection	358
12.3.2	Backward Elimination	358
12.3.3	Stepwise Regression	358
12.3.4	Sequential Variable-Selection with the SAT Data	359
12.3.5	Compounded Uncertainty in Stepwise Procedures	361

12.4	Model Selection Among All Subsets	361
12.4.1	The Cp Statistic and Cp Plot	362
12.4.2	Akaike and Bayesian Information Criteria	364
12.5	Posterior Beliefs About Different Models	365
12.6	Analysis of the Sex Discrimination Data	366
12.7	Related Issues	370
12.7.1	The Trouble with Interpreting Significance when Explanatory Variables Are Correlated	370
12.7.2	Regression for Adjustment and Ranking	371
12.7.3	Saturated Second-Order Models	372
12.7.4	Cross Validation	373
12.8	Summary	374
12.9	Exercises	375
	Conceptual Exercises	375
	Computational Exercises	376
	Data Problems	378
	Answers to Conceptual Exercises	382

CHAPTER 13 The Analysis of Variance for Two-Way Classifications 384

13.1	Case Studies	385
13.1.1	Intertidal Seaweed Grazers—A Randomized Experiment	385
13.1.2	The Pygmalion Effect—A Randomized Experiment	387
13.2	Additive and Nonadditive Models for Two-Way Tables	389
13.2.1	The Additive Model	389
13.2.2	The Saturated, Nonadditive Model	391
13.2.3	A Strategy for Analyzing Two-Way Tables with Several Observations per Cell	392
13.2.4	The Analysis of Variance <i>F</i> -Test for Additivity	392
13.3	Analysis of the Seaweed Grazer Data	393
13.3.1	Initial Assessment of Additivity, Outliers, and the Need for Transformation	393
13.3.2	The Analysis of Variance Table from the Fit to the Saturated Model	395
13.3.3	The Analysis of Variance Table for the Fit to the Additive Model	396
13.3.4	Answers to Specific Questions of Interest Using Linear Combinations	398
13.3.5	Answers to Specific Questions of Interest Using Multiple Regression with Indicator Variables	401
13.4	Analysis of the Pygmalion Data	402
13.4.1	Initial Exploration and Check on Additive Model	402
13.4.2	Answering the Question of Interest with Regression	405
13.4.3	A Closer Look at the Regression Estimate of Treatment Effect	405
13.4.4	The <i>p</i> -Value in the Randomization Distribution	407
13.5	Related Issues	408
13.5.1	Additivity and Nonadditivities	408
13.5.2	Orthogonal Contrasts	411
13.5.3	Randomized Blocks and Paired- <i>t</i> Analyses	411
13.5.4	Should Insignificant Block Effects Be Eliminated from the Model?	411
13.5.5	Multiple Comparisons	411
13.5.6	An Alternate Parameterization for the Additive Model	412
13.6	Summary	413

13.7	Exercises 414	
	Conceptual Exercises 414	
	Computational Exercises 415	
	Data Problems 417	
	Answers to Conceptual Exercises 419	
CHAPTER 14	Multifactor Studies Without Replication	420
14.1	Case Studies 421	
	14.1.1 Chimpanzees Learning Sign Language—A Controlled Experiment 421	
	14.1.2 Effects of Ozone in Conjunction with Sulfur Dioxide and Water Stress on Soybean Yield—A Randomized Experiment 422	
14.2	Strategies for Analyzing Tables with One Observation per Cell 425	
	14.2.1 Rationale for Designs with One Observation per Cell 425	
	14.2.2 Strategy for Data Analysis in the Absence of Replicates 426	
14.3	Analysis of the Chimpanzee Learning Times Study 427	
14.4	Analysis of the Soybean Data 432	
	14.4.1 Exploratory Analysis 433	
	14.4.2 Answering Questions of Interest with the Fitted Models 436	
14.5	Related Issues 439	
	14.5.1 Random Effects Models 439	
	14.5.2 Nested Classifications in the Analysis of Variance 439	
	14.5.3 Further Rationale for One Observation per Cell 440	
	14.5.4 Uniformity Trials 442	
14.6	Summary 443	
14.7	Exercises 443	
	Conceptual Exercises 443	
	Computational Exercises 444	
	Data Problems 445	
	Answers to Conceptual Exercises 448	
CHAPTER 15	Adjustment for Serial Correlation	449
15.1	Case Studies 450	
	15.1.1 Logging Practices and Water Quality—An Observational Study 450	
	15.1.2 Measuring Global Warming—An Observational Study 451	
15.2	Comparing the Means of Two Time Series 452	
	15.2.1 Serial Correlation and Its Effect on the Average of a Time Series 453	
	15.2.2 The Standard Error of an Average in a Serially Correlated Time Series 453	
	15.2.3 Estimating the First Serial Correlation Coefficient 456	
	15.2.4 Pooling Estimates and Comparing Means of Two Independent Time Series with the Same First Serial Correlation 457	
15.3	Regression After Transformation in the AR(1) Model 458	
	15.3.1 The Serial Correlation Coefficient Based on Regression Residuals 458	
	15.3.2 Regression with Filtered Variables 459	
15.4	Determining If Serial Correlation Is Present 461	
	15.4.1 An Easy, Large-Sample Test for Serial Correlation 461	
	15.4.2 The Nonparametric Runs Test 462	

15.5	Diagnostic Procedures for Judging the Adequacy of the AR(1) Model	463
15.5.1	When Is a Transformation of a Time Series Indicated?	464
15.5.2	The Partial Autocorrelation Function (PACF)	464
15.5.3	Bayesian Information Criterion	467
15.6	Related Issues	467
15.6.1	Time Series Analysis Is a Large-Sample Game	467
15.6.2	The Autocorrelation Function	468
15.6.3	Time Series Without a Time Series Package	468
15.6.4	Negative Serial Correlation	468
15.7	Summary	468
15.8	Exercises	469
	Conceptual Exercises	469
	Computational Exercises	470
	Data Problems	472
	Answers to Conceptual Exercises	475
CHAPTER 16 Repeated Measures and Other Multivariate Responses		476
16.1	Case Studies	477
16.1.1	Sites of Short- and Long-Term Memory—A Controlled Experiment	477
16.1.2	Oat Bran and Cholesterol—A Randomized Crossover Experiment	478
16.2	Tools and Strategies for Analyzing Repeated Measures	479
16.2.1	Types of Repeated Measures Studies	481
16.2.2	Profile Plots for Graphical Exploration	482
16.2.3	Strategies for Analyzing Repeated Measures	482
16.3	Comparing the Means of Bivariate Responses in Two Groups	485
16.3.1	Summary Statistics for Bivariate Responses	485
16.3.2	Pooled Variability Estimates	487
16.3.3	Hotelling's T^2 Statistic	488
16.3.4	Checking on Assumptions	490
16.3.5	Confidence Ellipses and Individual Confidence Intervals for Differences in Bivariate Means	491
16.4	One-Sample Analysis with Bivariate Responses	493
16.4.1	Treatment Differences in the Oat Bran Study	493
16.4.2	Summary Statistics for a Single Sample of Bivariate Responses	494
16.4.3	Hotelling's T^2 Test That the Means of a Bivariate Response Are Both Zero	494
16.4.4	Checking on Assumptions	497
16.5	Related Issues	497
16.5.1	Two-Sample Analysis with More Than Two Responses	497
16.5.2	One-Sample Analysis with More Than Two Responses	498
16.5.3	Multivariate Regression and Multivariate Analysis of Variance (MANOVA)	498
16.5.4	Planned and Unplanned Summaries of Multivariate Measurements as Response Variables	499
16.5.5	Planning an Experiment: Benefits of Repeated Measurements	499
16.5.6	Notes Concerning Correlation	500
16.5.7	Star Plots for Multivariate Responses	502
16.5.8	Controlling False Discovery Rate in Large Families of Hypothesis Tests	502
16.6	Summary	505

16.7	Exercises	506
	Conceptual Exercises	506
	Computational Exercises	507
	Data Problems	510
	Answers to Conceptual Exercises	513

CHAPTER 17 Exploratory Tools for Summarizing Multivariate Responses 514

17.1	Case Studies	515
	17.1.1 Magnetic Force on Rods in Printers—A Controlled Experiment	515
	17.1.2 Love and Marriage—An Observational Study	517
17.2	Linear Combinations of Variables	519
17.3	Principal Components Analysis	521
	17.3.1 The PCA Train	521
	17.3.2 Principal Components	521
	17.3.3 Variables Suggested by PCA	524
	17.3.4 Scatterplots in Principal Component Space	526
	17.3.5 The Factor Analysis Model and Principal Components Analysis	527
	17.3.6 PCA Usage	527
17.4	Canonical Correlations Analysis	528
	17.4.1 <i>Canonical Variables</i>	528
	17.4.2 Variables Suggested by CCA	530
	17.4.3 Love and Marriage Example	530
17.5	Introduction to Other Multivariate Tools	531
	17.5.1 Discriminant Function Analysis (DFA)	532
	17.5.2 Cluster Analysis	532
	17.5.3 Multidimensional Scaling	534
	17.5.4 Correspondence Analysis	534
	17.5.5 PCA and Empirical Orthogonal Functions (EOFs)	536
17.6	Summary	539
17.7	Exercises	541
	Conceptual Exercises	541
	Computational Exercises	543
	Data Problems	544
	Answers to Conceptual Exercises	548

CHAPTER 18 Comparisons of Proportions or Odds 549

18.1	Case Studies	550
	18.1.1 Obesity and Heart Disease—An Observational Study	550
	18.1.2 Vitamin C and the Common Cold—A Randomized Experiment	551
	18.1.3 Smoking and Lung Cancer—A Retrospective Observational Study	552
18.2	Inferences for the Difference of Two Proportions	553
	18.2.1 The Sampling Distribution of a Sample Proportion	553
	18.2.2 Sampling Distribution for the Difference Between Two Sample Proportions	555
	18.2.3 Inferences About the Difference Between Two Population Proportions	557

18.3	Inference About the Ratio of Two Odds	559
18.3.1	A Problem with the Difference Between Proportions	559
18.3.2	Odds	559
18.3.3	The Ratio of Two Odds	560
18.3.4	Sampling Distribution of the Log of the Estimated Odds Ratio	561
18.4	Inference from Retrospective Studies	563
18.4.1	Retrospective Studies	563
18.4.2	Why the Odds Ratio Is the Only Appropriate Parameter If the Sampling Is Retrospective	564
18.5	Summary	565
18.6	Exercises	566
	Conceptual Exercises	566
	Computational Exercises	568
	Data Problems	570
	Answers to Conceptual Exercises	571

CHAPTER 19 More Tools for Tables of Counts 572

19.1	Case Studies	573
19.1.1	Sex Role Stereotypes and Personnel Decisions—A Randomized Experiment	573
19.1.2	Death Penalty and Race of Murder Victim—An Observational Study	574
19.2	Population Models for 2×2 Tables of Counts	575
19.2.1	Hypotheses of Homogeneity and of Independence	575
19.2.2	Sampling Schemes Leading to 2×2 Tables	576
19.2.3	Testable Hypotheses and Estimable Parameters	578
19.3	The Chi-Squared Test	579
19.3.1	The Pearson Chi-Squared Test for Goodness of Fit	579
19.3.2	Chi-Squared Test of Independence in a 2×2 Table	580
19.3.3	Equivalence of Several Tests for 2×2 Tables	580
19.4	Fisher's Exact Test: The Randomization (Permutation) Test for 2×2 Tables	582
19.4.1	The Randomization Distribution of the Difference in Sample Proportions	582
19.4.2	The Hypergeometric Formula for One-Sided <i>p</i> -Values	584
19.4.3	Fisher's Exact Test for Observational Studies	584
19.4.4	Fisher's Exact Test Versus Other Tests	585
19.5	Combining Results from Several Tables with Equal Odds Ratios	586
19.5.1	The Mantel–Haenszel Excess	586
19.5.2	The Mantel–Haenszel Test for Equal Odds in Several 2×2 Tables	588
19.5.3	Estimate of the Common Odds Ratio	589
19.6	Related Issues	590
19.6.1	$r \times c$ Tables of Counts	590
19.6.2	Higher-Dimensional Tables of Counts	591
19.7	Summary	591
19.8	Exercises	592
	Conceptual Exercises	592
	Computational Exercises	594
	Data Problems	596
	Answers to Conceptual Exercises	598

CHAPTER 20	Logistic Regression for Binary Response Variables	601
20.1	Case Studies 602	
20.1.1	Survival in the Donner Party—An Observational Study 602	
20.1.2	Birdkeeping and Lung Cancer—A Retrospective Observational Study 602	
20.2	The Logistic Regression Model 604	
20.2.1	Logistic Regression as a Generalized Linear Model 606	
20.2.2	Interpretation of Coefficients 608	
20.3	Estimation of Logistic Regression Coefficients 609	
20.3.1	Maximum Likelihood Parameter Estimation 609	
20.3.2	Tests and Confidence Intervals for Single Coefficients 612	
20.4	The Drop-in-Deviance Test 614	
20.5	Strategies for Data Analysis Using Logistic Regression 617	
20.5.1	Exploratory Analysis 617	
20.6	Analyses of Case Studies 618	
20.6.1	Analysis of Donner Party Data 618	
20.6.2	Analysis of Birdkeeping and Lung Cancer Data 619	
20.7	Related Issues 622	
20.7.1	Matched Case–Control Studies 622	
20.7.2	Probit Regression 623	
20.7.3	Discriminant Analysis Using Logistic Regression 623	
20.7.4	Monte Carlo Methods 623	
20.8	Summary 625	
20.9	Exercises 626	
	Conceptual Exercises 626	
	Computational Exercises 627	
	Data Problems 629	
	Answers to Conceptual Exercises 632	
CHAPTER 21	Logistic Regression for Binomial Counts	634
21.1	Case Studies 635	
21.1.1	Island Size and Bird Extinctions—An Observational Study 635	
21.1.2	Moth Coloration and Natural Selection—A Randomized Experiment 637	
21.2	Logistic Regression for Binomial Responses 639	
21.2.1	Binomial Responses 639	
21.2.2	The Logistic Regression Model for Binomial Responses 640	
21.3	Model Assessment 640	
21.3.1	Scatterplot of Empirical Logits Versus an Explanatory Variable 640	
21.3.2	Examination of Residuals 641	
21.3.3	The Deviance Goodness-of-Fit Test 642	
21.4	Inferences About Logistic Regression Coefficients 644	
21.4.1	Wald’s Tests and Confidence Intervals for Single Coefficients 644	
21.4.2	The Drop-in-Deviance Test 645	
21.5	Extra-Binomial Variation 646	
21.5.1	Extra-Binomial Variation and the Logistic Regression Model 647	
21.5.2	Checking for Extra-Binomial Variation 647	
21.5.3	Quasi-Likelihood Inference when Extra-Binomial Variation Is Present 648	
21.6	Analysis of Moth Predation Data 650	

21.7	Related Issues 654	
21.7.1	Why the Deviance Changes when Binary Observations Are Grouped 654	
21.7.2	Logistic Models for Multilevel Categorical Responses 655	
21.7.3	The Maximum Likelihood Principle 656	
21.7.4	Bayesian Inference 660	
21.7.5	Generalized Estimating Equations 663	
21.8	Summary 664	
21.9	Exercises 665	
	Conceptual Exercises 665	
	Computational Exercises 666	
	Data Problems 667	
	Answers to Conceptual Exercises 671	
CHAPTER 22	Log-Linear Regression for Poisson Counts	672
22.1	Case Studies 673	
22.1.1	Age and Mating Success of Male Elephants—An Observational Study 673	
22.1.2	Characteristics Associated with Salamander Habitat 674	
22.2	Log-Linear Regression for Poisson Responses 676	
22.2.1	Poisson Responses 676	
22.2.2	The Poisson Log-Linear Model 677	
22.2.3	Estimation by Maximum Likelihood 679	
22.3	Model Assessment 679	
22.3.1	Scatterplot of Logged Counts Versus an Explanatory Variable 679	
22.3.2	Residuals 679	
22.3.3	The Deviance Goodness-of-Fit Test 680	
22.3.4	The Pearson Chi-Squared Goodness-of-Fit Test 682	
22.4	Inferences About Log-Linear Regression Coefficients 682	
22.4.1	Wald's Test and Confidence Interval for Single Coefficients 683	
22.4.2	The Drop-in-Deviance Test 683	
22.5	Extra-Poisson Variation and the Log-Linear Model 684	
22.5.1	Extra-Poisson Variation 684	
22.5.2	Checking for Extra-Poisson Variation 685	
22.5.3	Inferences when Extra-Poisson Variation Is Present 686	
22.6	Related Issues 688	
22.6.1	Log-Linear Models for Testing Independence in Tables of Counts 688	
22.6.2	Poisson Counts from Varying Effort 690	
22.6.3	Negative Binomial Regression 691	
22.7	Summary 692	
22.8	Exercises 693	
	Conceptual Exercises 693	
	Computational Exercises 694	
	Data Problems 695	
	Answers to Conceptual Exercises 697	
CHAPTER 23	Elements of Research Design	699
23.1	Case Study 700	
23.1.1	Biological Control of a Noxious Weed—A Randomized Experiment 700	
23.2	Considerations in Forming Research Objectives 700	

23.3	Research Design Tool Kit 701
23.3.1	Controls and Placebos 701
23.3.2	Blinding 702
23.3.3	Blocking 702
23.3.4	Stratification 703
23.3.5	Covariates 703
23.3.6	Randomization 703
23.3.7	Random Sampling 704
23.3.8	Replication 704
23.3.9	Balance 704
23.4	Design Choices That Affect Accuracy and Precision 705
23.4.1	Attaching Desired Precision to Practical Significance 705
23.4.2	How to Improve a Confidence Interval 706
23.5	Choosing a Sample Size 708
23.5.1	Studies with a Numerical Response 708
23.5.2	Studies Comparing Two Proportions 709
23.5.3	Sample Size for Estimating a Regression Coefficient 710
23.6	Steps in Designing a Study 711
23.6.1	Stating the Objective 711
23.6.2	Determining the Scope of Inference 712
23.6.3	Understanding the System 713
23.6.4	Deciding How to Measure a Response 713
23.6.5	Listing Factors That Can Affect the Response 714
23.6.6	Planning the Conduct of the Experiment 715
23.6.7	Outlining the Statistical Analysis 716
23.6.8	Determining the Sample Size 716
23.7	Related Issue—A Factor of 4 719
23.8	Summary 720
23.9	Exercises 720
	Conceptual Exercises 720
	Study Designs for Discussion 721
	Computational Exercises 722
	Design Problems 723
	Answers to Conceptual Exercises 723

CHAPTER 24 Factorial Treatment Arrangements and Blocking Designs 725

24.1	Case Study 726
24.1.1	Amphibian Crisis Linked to Ultraviolet—A Randomized Experiment 726
24.2	Treatments 727
24.2.1	Choosing Treatment Levels 727
24.2.2	The Rationale for Several Factors 728
24.3	Factorial Arrangement of Treatment Levels 729
24.3.1	Definition and Terminology for a Factorial Arrangement 729
24.3.2	The 2^2 Factorial Structure 729
24.3.3	The 2^3 Factorial Structure 732
24.3.4	The 3^2 Factorial Structure 734
24.3.5	Higher-Order Factorial Arrangements 737

24.4	Blocking	739
24.4.1	Randomized Blocks	739
24.4.2	Latin Square Blocking	742
24.4.3	Split-Plot Designs	742
24.5	Summary	743
24.6	Exercises	744
	Conceptual Exercises	744
	Computational Exercises	745
	Data Problems	746
	Answers to Conceptual Exercises	747

APPENDIX	Bibliography	748
-----------------	---------------------	------------

Index	751
--------------	------------



Preface

This book is written for those who need to use statistical methods to analyze data from experiments and observational studies and who need to communicate the results to others. It is intended as a text for graduate students who are preparing to design, implement, analyze, and report their research. The students must have some knowledge of basic statistical concepts such as means, standard deviations, histograms, the normal and t -distributions, but they need not be familiar with calculus or matrix algebra. All should have access to a statistical software package and a moderately powerful computer.

TO THE STUDENT

Statistics is like grout—the word feels decidedly unpleasant in the mouth, but it describes something essential for holding a mosaic in place. Statistics is a common bond supporting all other sciences. It provides standards of empirical proof and a language for communicating scientific results. Statistical sleuthing is the process of using statistical tools to answer questions of interest. It includes devising experiments to unearth hidden truths, describing real data using tools based on ideal mathematical models, answering the questions of interest efficiently, verifying that the tools are appropriate, and snooping around to see if there is anything more to be learned. *The Statistical Sleuth* will show you how this all comes about.

Case Studies

The Statistical Sleuth is organized around case studies, which begin each chapter and are used throughout to illustrate how the statistical tools operate. A small section entitled *Statistical Conclusions* accompanies each case study, demonstrating how to communicate statistical findings for a research publication. You should realize that the methods upon which the findings are based will be foreign to you upon first reading. After the chapter has been read, you should return to the studies and consider carefully how the chapter's methods have been used to answer the questions posed by the researchers.

Examine each case study carefully for its structural design. Ask yourself why the study was structured in the way it was. The studies will not only illustrate analytical techniques; most also present exemplary structures for your own studies.

Mathematical Level

The emphasis of this book is on the practical use of statistical methods. The correct practical use of statistical tools requires some understanding of the mathematical foundation for the tools. Sometimes algebra or elementary mathematical statistics are the best device for communicating the motivation. In general, however, the level of mathematics required to follow this book is not high.

What will you learn?

Do not expect to learn all that you will need to make you an experienced analyst. You will improve your understanding of statistical reasoning and of measures of uncertainty. You will learn how to translate mountains of computer output into short summary statements that communicate the results in a language common to all scientists. You will also learn a fairly large array of statistical tools that will be useful for a wide range of problems. But there are many more tools that are not covered in this book and many lessons that can only be gained through experience. At some point you may need to seek the help of a professional statistician. Then, at least, you will know the language, the general tools, and the spirit of statistical data analysis, which will make communication with a statistician more effective and beneficial.

Resources

Visit the Sleuth Web page or access course materials and companion resources at www.cengagebrain.com. At the cengagebrain.com home page, search for the ISBN of your title (from the back of your book) using the search box at the top of the page. This will take you to the product page where free companion resources can be found. Or go to www.StatisticalSleuth.com for answers to selected exercises, updates, and links to other sites: <http://www.statisticalsleuth.com>.

TO THE INSTRUCTOR

Level of Sophistication

The level of sophistication for this text is high when it comes to models and methods needed to analyze data and interpret results, but low when it comes to mathematics. Our foremost concern is that future researchers learn proper approaches for conducting the statistical aspects of their research. To this end, mathematics is neither sought out nor avoided.

Case Studies

Most chapters begin with two case studies for motivation and demonstration. Making these studies a central feature forces us to consider applied statistics more seriously than if we simply provided a data set to demonstrate a particular tool. It compels us to maintain a question-driven approach to the analysis of data.

The case studies are also our tool for exciting students. We cannot successfully teach them if they are not genuinely interested. We have tried to find a variety of interesting real data sets where the statistical analysis provides useful answers to

questions of interest. In some cases, we found descriptions and summary statistics that made excellent examples, but we were unable to obtain the raw data. We have still used some of them, however, by generating data to match the summary statistics. To identify these cases, we use the phrase “based on a real study.”

Although we have made the data problems central, limitations of space prevent us from including all the graphical displays and the different analyses we would like to present. We encourage the instructor to go into more depth in showing computer output, graphical displays, and alternative analyses.

The Starting Point

At first glance, the first four chapters of *The Statistical Sleuth* appear to rehash the topics of one-sample and two-sample analysis, which are covered in introductory courses. This is not the case. These chapters are intended as a model for how topics in the rest of the book are treated. The chapters provide a detailed examination of material to which students have already been exposed, and an introduction to a philosophy of learning and using statistics.

Material Covered

The Statistical Sleuth's principal tool is regression analysis. We have added several topics that are not ordinarily covered in a regression text: (1) *Generalized linear models*, including logistic and log-linear regression. This important tool enables researchers to analyze a wide range of problems that have until recently been analyzed with inappropriate tools (ANOVA) or with appropriate but difficult tools (contingency table chi-squares). We stress the parallels between generalized linear model regression and ordinary regression. With calculations provided by statistical software, this tool has become entirely understandable and extremely valuable. (2) *Repeated measures*. Whereas there is a great tendency for researchers to turn to a statistical computer packaged function that has “repeated measures” in the title, we feel there needs to be more guidance on a strategy for considering such data analysis. Chapters 16 and 17 respectively emphasize question-driven and data-driven reduction of dimensionality. (3) *Serial correlation*. Although a full treatment of time series analysis is beyond the scope of this book, by adjusting and filtering for the first-order autoregression we provide tools that expand the usefulness of regression technology to problems involving serial correlation.

Our decisions regarding coverage reflect the kinds of problems graduate researchers typically encounter. The topics chosen arise repeatedly in the campus-wide consulting service operated by Oregon State University's Department of Statistics for faculty and graduate students. By offering a textbook with these topics, we hope to relieve the pressure on departments to offer separate courses in categorical data analysis, in multivariate analysis, and in time series analysis for nonstatistics majors, or to enroll nonstatistics majors in classes designed primarily for statistics majors.

Possible Paths Through the Chapters

The Statistical Sleuth was designed for a three-quarter sequence covering eight chapters each term. Typically not all the material is covered, and we have provided

a number of optional topics in a section titled *Related Issues* at the end of most chapters. The book may also be used for a two-semester class in its entirety. For a one-semester or two-quarter class, we recommend the following sequence: conclusions and interpretations (1–4), several sample problems (5–6), simple linear regression (7–8), multiple regression (9–12), two-way analysis of variance (13–14), and logistic regression (20–21). There is room to mix and match to meet specific needs.

The Statistical Sleuth covers regression prior to two-way analysis of variance, in contrast with the more traditional presentation of two-way ANOVA directly after one-way ANOVA. Our reasoning here is that regression tools applied to the two-way situation are easier to interpret. They are also less subject to misunderstandings arising from imbalance in the experimental or sampling design. Experimental design chapters (23–24) appear at the end of the book. In truth, design issues are discussed throughout the text as they apply to the case studies. Topics such as replication, blocking, factorial treatment structures, and randomization appear repeatedly. So the actual chapters on design organize and summarize the issues. We also believe it is difficult to design an experiment without an understanding of the applicable analytic tools.

Web Site

To access additional course materials and companion resources, please visit www.cengagebrain.com. At the cengagebrain.com home page, search for the ISBN of your title (from the back cover of your book) using the search box at the top of the page. This will take you to the product page where free companion resources can be found. Or go to www.StatisticalSleuth.com.

Statistical Computer Programs

A computer and a packaged statistical computer program are essential companions for *The Statistical Sleuth*. Many good packages are available. Unfortunately, they differ considerably in their style, language, and output. Some instruction about the particular software package must accompany your instruction for using the statistical tools in this book. The student's statistical analysis, however, should be guided by good statistical strategy, and not by the package, which is just a means for accomplishing the end. The data sets presented as case studies and as exercises are available online at www.cengagebrain.com.

ACKNOWLEDGMENTS

We take pleasure in expressing our gratitude to those who have helped us with this project. We are especially indebted to those who taught from earlier versions of this book. Helen Berg (mayor of Corvallis), Ginny Lesser (Oregon State University), Loretta Thielman (University of Wisconsin—Stout), and Peter Thielman (University of Wisconsin—Stout) all cheerfully endured mistakes and made valuable suggestions for improvements. Thanks also to other colleagues who read drafts and provided suggestions, including Rick Rossi, Mike Scott, and Jeannie Sifneos.

We are very grateful for the use of the facilities at the Department of Statistics at Oregon State University, and the supportive cooperation of the chairman, Jus-tus Seely, and the assistance of Genevieve Downing. Our views of statistics have benefited from fruitful discussions with our colleagues in the Statistics Department.

A portion of the early work on the book was conducted while Dan Schafer was on sabbatical leave at The University of Western Australia. The Mathematics Department there supplied facilities for which we are grateful. Special thanks go to Ian James, currently at Murdoch University in Perth.

A very enjoyable aspect of this project for us has been the discovery of modern scientific applications we have found while searching for case studies. We have gained inspiration from the many scientists who know good questions to ask in their research and whose scientific creativity is spurred by a genuine interest in finding the answers.

Very special thanks go to the students of Statistics 511–513 classes at Oregon State University. Their maturity and collective knowledge about scientific subjects have made the course a pleasure to teach. We have been driven by a desire to provide them with tools and strategies they may use to become the scientists that inspire us in the future.

We were guided through production by the staff at Hearthside Publishing Services, Anne Seitz and Laura Horowitz.

Finally, we wish to express our gratitude to the following reviewers: Andrew Barron, Yale University; Darl Bien, University of Denver; Phil Everson, Swarthmore University; Mauro Gasparini, Purdue University; Joseph Glaz, University of Connecticut; Mark Kaiser, Iowa State University; Frank Martin, University of Minnesota; Michael Martin, Australian National University; David Mathias, Rochester Institute of Technology; Julia Norton, California State University—Hayward; John Rawlings, North Carolina State University; Larry Ringer, Texas A&M University; Bruce Schaalje, Brigham Young University; Joseph Schafer, Pennsylvania State University; Kirk Steinhorst, University of Idaho; Loretta Thielman, University of Wisconsin—Stout; Bruce Trumbo, California State University—Hayward; and John Wasik, North Carolina State University.

Third Edition Notes

We dropped the statistical tables at the end of the book—it was time for them to go. We've added approximately 70 new data problems for exercises, as well as new sections on the Dunnett multiple comparison procedure, reasoning fallacies associated with statistical hypothesis testing, control of false discovery rates for large families of hypothesis tests, Monte Carlo methods, negative binomial regression, and generalized estimating equations. We are grateful to the following individuals for helpful comments and suggestions used in this revision: Scott Freeman, Eugene Gallagher, Alix Gitelman, Megan Higgs, Elizabeth Housworth, Martin Jones, William Kitto, Tim Leonard, Virginia Lesser, Lisa Madsen, Xiao Li Meng, Paul Murtaugh, Roger Pinkham, Berwin Turlach, Edward Whitney, and Captain William M. Wilder.

*Fred Ramsey
Daniel Schafer*

Drawing Statistical Conclusions

Statistical sleuthing means carefully examining data to answer questions of interest. This book is about the process of statistical sleuthing, including strategies and tools for answering questions of interest and guidelines for interpreting and communicating results.

This chapter is about interpreting statistical results—a process crucially linked to study design. The setting for this and the next three chapters is the two-sample problem, where it is convenient to illustrate concepts and strategies employed in all statistical analysis.

An answer to a question is accompanied by a statistical measure of uncertainty, which is based on a probability model. When that probability model is based on a chance mechanism—like the flipping of coins to decide which subjects get which treatment or the drawing of lottery tickets to decide which members of a population get selected in a sample—measures of uncertainty and the inferences drawn from them are formally justified. Often, however, chance mechanisms are invented as conceptual frameworks for drawing statistical conclusions. For understanding and communicating statistical conclusions it is important to understand whether a chance mechanism was used and, if so, whether it was used for sample selection, group allocation, or both.

1.1 CASE STUDIES

1.1.1 Motivation and Creativity—A Randomized Experiment

Do grading systems promote creativity in students? Do ranking systems and incentive awards increase productivity among employees? Do rewards and praise stimulate children to learn? Although reward systems are deeply embedded in schools and in the workplace, a growing body of evidence suggests that rewards may operate in precisely the opposite way from what is intended.

A remarkable demonstration of this phenomenon was provided by psychologist Teresa Amabile in an experiment concerning the effects of intrinsic and extrinsic motivation on creativity. Subjects with considerable experience in creative writing were randomly assigned to one of two treatment groups: 24 of the subjects were placed in the “intrinsic” treatment group, and 23 in the “extrinsic” treatment group, as indicated in Display 1.1. The “intrinsic” group completed the questionnaire at the top of Display 1.2. The questionnaire, which involved ranking intrinsic reasons for writing, was intended as a device to establish a thought pattern concerning intrinsic motivation—doing something because doing it brings satisfaction. The “extrinsic” group completed the questionnaire at the bottom of Display 1.2, which was used as a device to get this group thinking about extrinsic motivation—doing something because a reward is associated with its completion.

DISPLAY 1.1

Creativity scores in two motivation groups, and their summary statistics

	Intrinsic group		Extrinsic group	
	12.0	20.5	5.0	17.4
	12.0	20.6	5.4	17.5
	12.9	21.3	6.1	18.5
	13.6	21.6	10.9	18.7
	16.6	22.1	11.8	18.7
	17.2	22.2	12.0	19.2
	17.5	22.6	12.3	19.5
	18.2	23.1	14.8	20.7
	19.1	24.0	15.0	21.2
	19.3	24.3	16.8	22.1
	19.8	26.7	17.2	24.0
	20.3	29.7	17.2	
Sample Size:	24		23	
Average:	19.88		15.74	
Sample Standard Deviation:	4.44		5.25	

After completing the questionnaire, all subjects were asked to write a poem in the Haiku style about “laughter.” All poems were submitted to 12 poets, who evaluated them on a 40-point scale of creativity, based on their own subjective views. Judges were not told about the study’s purpose. The average ratings given by the 12 judges are shown for each of the study subjects in Display 1.1. (Data based

DISPLAY 1.2

Questionnaires given creative writers, to rank intrinsic and extrinsic reasons for writing

INSTRUCTIONS: Please rank the following list of reasons for writing, in order of personal importance to you (1 = highest, 7 = lowest).

- You get a lot of pleasure out of reading something good that you have written.
- You enjoy the opportunity for self-expression.
- You achieve new insights through your writing.
- You derive satisfaction from expressing yourself clearly and eloquently.
- You feel relaxed when writing.
- You like to play with words.
- You enjoy becoming involved with ideas, characters, events, and images in your writing.

*List of extrinsic
reasons for writing.*

*List of intrinsic
reasons for writing.*

INSTRUCTIONS: Please rank the following list of reasons for writing, in order of personal importance to you (1 = highest, 7 = lowest).

- You realize that, with the introduction of dozens of magazines every year, the market for free-lance writing is constantly expanding.
- You want your writing teachers to be favorably impressed with your writing talent.
- You have heard of cases where one best-selling novel or collection of poems has made the author financially secure.
- You enjoy public recognition of your work.
- You know that many of the best jobs available require good writing skills.
- You know that writing ability is one of the major criteria for acceptance into graduate school.
- Your teachers and parents have encouraged you to go into writing.

on the study in T. Amabile, “Motivation and Creativity: Effects of Motivational Orientation on Creative Writers,” *Journal of Personality and Social Psychology* 48(2) (1985): 393–99.) Is there any evidence that creativity scores tend to be affected by the type of motivation (intrinsic or extrinsic) induced by the questionnaires?

Statistical Conclusion

This experiment provides strong evidence that receiving the “intrinsic” rather than the “extrinsic” questionnaire caused students in this study to score higher on poem creativity (two-sided p -value = 0.005 from a two-sample t -test as an approximation to a randomization test). The estimated treatment effect—the increase in score attributed to the “intrinsic” questionnaire—is 4.1 points (95% confidence interval: 1.3 to 7.0 points) on a 0–40-point scale.

Note: The statistical conclusions associated with the case studies in each chapter use tools and, in some cases, terms that are introduced later. You should read the conclusions initially to appreciate the scientific statements that can be

drawn from the chapter's tools, without worrying about the parts you don't understand, and then return to them again after reading the chapter for a more complete understanding.

Scope of Inference

Since this was a randomized experiment, one may infer that the difference in creativity scores was *caused* by the difference in motivational questionnaires. Because the subjects were not selected randomly from any population, extending this inference to any other group is speculative. This deficiency, however, is minor; the causal conclusion is strong even if it applies only to the recruited subjects.

1.1.2 Sex Discrimination in Employment—An Observational Study

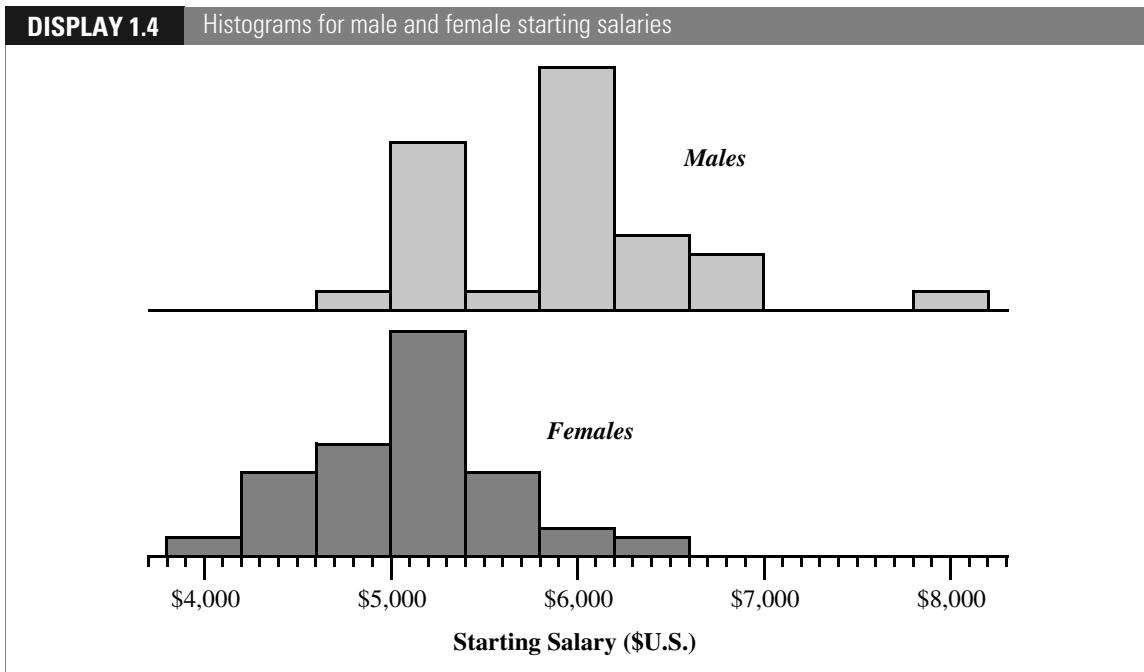
Did a bank discriminatorily pay higher starting salaries to men than to women? The data in Display 1.3 are the beginning salaries for all 32 male and all 61 female skilled, entry-level clerical employees hired by the bank between 1969 and 1977. (Data from a file made public by the defense and described by H. V. Roberts, "Harris Trust and Savings Bank: An Analysis of Employee Compensation" (1979), Report 7946, Center for Mathematical Studies in Business and Economics, University of Chicago Graduate School of Business.)

DISPLAY 1.3 Starting salaries (\$U.S.) for 32 male and 61 female clerical hires at a bank

Males			Females					
4,620	5,700	6,000	3,900	4,500	4,800	5,220	5,400	5,640
5,040	6,000	6,000	4,020	4,620	4,800	5,220	5,400	5,700
5,100	6,000	6,000	4,290	4,800	4,980	5,280	5,400	5,700
5,100	6,000	6,300	4,380	4,800	5,100	5,280	5,400	5,700
5,220	6,000	6,600	4,380	4,800	5,100	5,280	5,400	5,700
5,400	6,000	6,600	4,380	4,800	5,100	5,400	5,400	5,700
5,400	6,000	6,600	4,380	4,800	5,100	5,400	5,400	6,000
5,400	6,000	6,840	4,380	4,800	5,100	5,400	5,520	6,000
5,400	6,000	6,900	4,440	4,800	5,100	5,400	5,520	6,120
5,400	6,000	6,900	4,500	4,800	5,160	5,400	5,580	6,300
		6,000						6,300
		8,100						

Statistical Conclusion

As evident in Display 1.4, the data provide convincing evidence that the male mean is larger than the female mean (one-sided p -value < 0.00001 from a two-sample t -test). The male mean exceeds the female mean by an estimated \$818 (95% confidence interval: \$560 to \$1,076).



Scope of Inference

Although there is convincing evidence that the males, as a group, received larger starting salaries than the females, the statistics alone cannot address whether this difference is attributable to sex discrimination. The evidence is consistent with discrimination, but other possible explanations cannot be ruled out; for example, the males may have had more years of previous experience.

1.2 STATISTICAL INFERENCE AND STUDY DESIGN

The inferences one may draw from any study depend crucially on the study's design. Two distinct forms of inference—causal inference and inference to populations—can be justified by the proper use of random mechanisms.

1.2.1 Causal Inference

In a *randomized experiment*, the investigator controls the assignment of experimental units to groups and uses a chance mechanism (like the flip of a coin) to make the assignment. The motivation and creativity study is a randomized experiment, because a random mechanism was used to assign the study subjects to the two motivation groups. In an *observational study*, the group status of the subjects is established beyond the control of the investigator. The sex discrimination study is

observational because the group status (sex of the employee) was, obviously, not decided by the investigator.

Causal Conclusions and Confounding Variables

Can statistical analysis alone be used to establish causal relationships? The answer is simple and concise:

Statistical inferences of cause-and-effect relationships can be drawn from randomized experiments, but not from observational studies.

In a rough sense, randomization ensures that subjects with different, and possibly relevant, features are mixed up between the two groups. Surely some of the subjects in the motivation and creativity experiment were naturally more creative than others. In view of the randomization, however, there is no reason to expect that they would be placed disproportionately in the intrinsic motivation group, since every subject had the same chance of being placed in that group. There is, of course, a chance that the randomization—the particular result of the random assignment—turned out by chance in such a way that the intrinsic motivation group received many more of the naturally creative writers. *This chance, however, is incorporated into the statistical tools that are used to express uncertainty.*

In an observational study, it is impossible to draw a causal conclusion from the statistical analysis alone. One cannot rule out the possibility that confounding variables are responsible for group differences in the measured outcome.

A confounding variable is related both to group membership and to the outcome. Its presence makes it hard to establish the outcome as being a direct consequence of group membership.

In fact, the males in the sex discrimination example generally did have more years of education than the females; and this, not sex, may have been responsible for the disparity in starting salaries. Thus, the effect of sex cannot be separated from—it is confounded with—the effect of education. Tools exist for comparing male and female salaries after accounting for the effect of education (see Chapter 9), and these help clarify matters to some extent. Other confounding variables, however, may not be recognized or measured, and these consequently cannot be accounted for in the analysis. Therefore, it may be possible to conclude that males tend to get larger starting salaries than females, even after accounting for years of education, and yet still not be possible to conclude, from the statistics alone, that this happens *because* they are males.

Do Observational Studies Have Value?

Is there any role at all for observational data in serious scientific inquiry? The following points indicate that the answer is yes.

1. *Establishing causation is not always the goal.* A study was conducted on 10 American men of Chinese descent and 10 American men of European descent to examine the effect of a blood-pressure-reducing drug. The result—that the men of Chinese ancestry tended to exhibit a different response to the drug—does not prove that being of Chinese descent is responsible for the difference. Diets, for example, may differ in the two groups and may be responsible for the different sensitivities. Nevertheless the study does provide important information for doctors prescribing the drug to people from these populations.
2. *Establishing causation may be done in other ways.* Although statistical methods cannot eliminate the possibility of confounding factors, there may be strong theoretical reasons for doing so. Radiation biologists counted chromosomal aberrations in a sample of Japanese atomic bomb survivors who received radiation from the blast, and compared these to counts on individuals who were far enough from the blast not to have received any radiation. Although the data are observational, the researchers are certain that higher counts in the radiation group can only be due to the radiation, and the data can therefore be used to estimate the dose-response relationship between radiation and chromosomal aberration.
3. *Analysis of observational data may lend evidence toward causal theories and suggest the direction of future research.* Many observational studies indicated an association between smoking and lung cancer, but causation was accepted only after decades of observational studies, experimental studies on laboratory animals, and a scientific theory for the carcinogenic mechanism of smoking.

Although observational studies are undoubtedly useful, the inappropriate practice of claiming or implying cause-and-effect relationships from them—sometimes in subtle ways—is widespread.

1.2.2 Inference to Populations

A second distinction between study designs relates to how units are selected. In a *random sampling* study, units are selected by the investigator from a well-defined population. All units in the population have a chance of being selected, and the investigator employs a chance mechanism (like a lottery) to determine actual selection. Neither of the case studies in Section 1.1 used random sampling. In contrast, the study units often are *self-selected*. The subjects of the creativity study volunteered their participation. The decision to volunteer can have a strong relationship with the outcome (creativity score), and this precludes the subjects from representing a broader population.

Again, the inferential situation is straightforward:

Inferences to populations can be drawn from random sampling studies, but not otherwise.

Random sampling ensures that all subpopulations are represented in the sample in roughly the same mix as in the overall population. Again, random selection

has a chance of producing nonrepresentative samples, but *the statistical inference procedures incorporate measures of uncertainty that describe that chance.*

Simple Random Sampling

The most basic form of random sampling is simple random sampling.

A simple random sample of size n from a population is a subset of the population consisting of n members selected in such a way that every subset of size n is afforded the same chance of being selected.

A typical method assigns each member of the population a computer-generated random number. When both lists are re-ordered according to increasing values of the random numbers, the sample of population members appears at the top of the list. Most statistical programs can do this automatically with a “sample” command.

1.2.3 Statistical Inference and Chance Mechanisms

Statistical analysis is used to make statements from available data in answer to questions of interest about some broader context than the study at hand. No such statement about the broader context can be made with absolute certainty, so every statement includes some measure of uncertainty. In statistical analysis, uncertainty is explained by means of chance models, which relate the available data to the broader context. These ideas are expressed in the following definitions:

An inference is a conclusion that patterns in the data are present in some broader context.
A statistical inference is an inference justified by a probability model linking the data to the broader context.

The probability models are associated with the chance mechanisms used to select units from a population or to assign units to treatment groups. They enable the investigator to calculate measures of uncertainty to accompany inferential conclusions.

When a chance mechanism has been used in the study, uncertainty measures accompanying the researcher’s inferences are backed by a *bona fide* probability structure that exactly describes the study. Often, however, units do not arise from random sampling of real populations; instead, the available units are self-selected or are the result of a haphazard selection procedure. For randomized experiments, this may be no problem, since cause-and-effect conclusions can be drawn regarding the effects on the particular units selected (as in the motivation and creativity study). These conclusions can be quite strong, even if the observed pattern cannot be inferred to hold in some general population. For observational studies, the lack of truly random samples is more worrisome, because making an inference about some larger population is usually the goal. It may be possible to *pretend* that the

units are as representative as a random sample, but the potential for bias from haphazard or convenience selection remains a serious concern.

Example

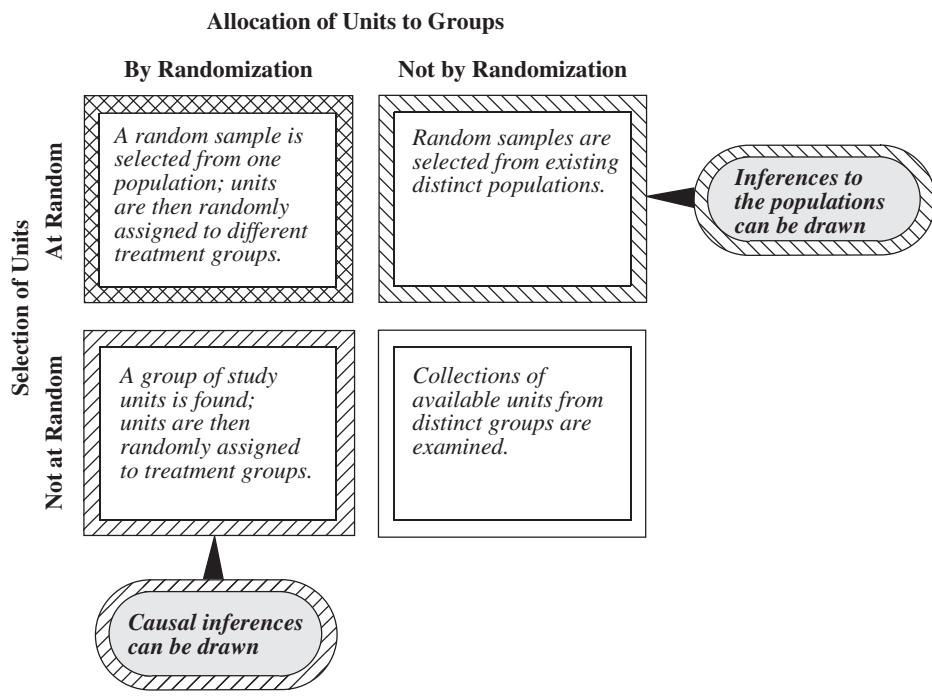
Researchers measured the lead content in teeth and the intelligence quotient (IQ) test scores for all 3,229 children attending first and second grades in the period between 1975 and 1978 in Chelsea and Somerville, Massachusetts. The IQ scores for those with low lead concentrations were found to be significantly higher than for those with high lead concentrations. What can be inferred? There is no random sample. In a strict sense, the statistical results apply only to the children actually measured; any extrapolation of the pattern to other children comes from supposing that the relationship between lead and IQ is similar for others. This is not necessarily a bad assumption. The point is that extending the inference to other children is surely open to question.

Statistical Inferences Based on Chance Mechanisms

Four situations are exhibited in Display 1.5, along with the types of inference that may be drawn in each. In observational studies, obtaining random samples from the populations of interest is often impractical or impossible, and inference based on

DISPLAY 1.5

Statistical inferences permitted by study designs



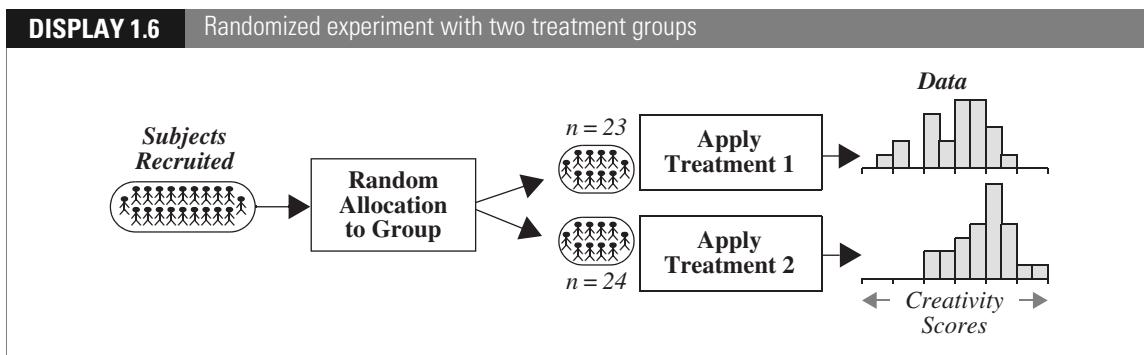
assumed models may be better than no inference at all. In controlled experiments, however, there is no excuse for avoiding randomization.

1.3 MEASURING UNCERTAINTY IN RANDOMIZED EXPERIMENTS

The critical element in understanding statistical measures of uncertainty is visualizing hypothetical replications of a study under different realizations of the chance mechanism used for sample selection or group assignment. This section describes the probability model for randomized experiments and illustrates how a measure of uncertainty is calculated to express the evidence for a treatment difference in the creativity study.

1.3.1 A Probability Model for Randomized Experiments

A schematic diagram for the creativity study (Display 1.6) is typical of a randomized experiment. The chance mechanism for randomizing units to treatment groups ensures that every subset of 24 subjects gets the same chance of becoming the intrinsic group. For example, 23 red and 24 black cards could be shuffled and dealt, one to each subject. Subjects are treated according to their assigned group, and the different treatments change their responses.



An Additive Treatment Effect Model

Let Y denote the creativity score that a subject would receive after exposure to the extrinsic questionnaire. A general model for the experiment would assert that this same subject would receive a different creativity score, Y^* , after exposure to the intrinsic questionnaire. An *additive treatment effect* model postulates that $Y^* = Y + \delta$. The treatment effect, δ , is a *parameter*—an unknown constant that describes a key feature in the model for answering questions of interest.

1.3.2 A Test for Treatment Effect in the Creativity Study

This section illustrates a statistical inference arising from the additive treatment effect model for the creativity study. The question concerns whether the treatment effect is real. The *p-value* will be introduced as a measure of uncertainty associated with the answer.

Null and Alternative Hypotheses

Questions of interest are translated into questions about parameters in probability models. This usually involves some simplification—such as the assumption that the intrinsic questionnaire adds the same amount to anyone’s creativity score—but the expression of a question in terms of a single parameter crystallizes the situation into one where powerful statistical tools can be applied.

In the creativity study, the question “Is there a treatment effect?” becomes a question about whether the parameter δ has a nonzero value. The statement that $\delta = 0$ is called the *null hypothesis*, while a proposition that $\delta \neq 0$ is an *alternative hypothesis*. In general, the null hypothesis is the one that specifies a simpler state of affairs; typically—as in this case—an absence of an effect.

A Test Statistic

A *statistic* is a numerical quantity calculated from data, like the average creativity score in the intrinsic group. A *test statistic* is a statistic used to measure the plausibility of an alternative hypothesis relative to a null hypothesis. For the hypothesis that δ equals zero, the test statistic is the difference in averages, $\bar{Y}_2 - \bar{Y}_1$. Its value should be “close to” zero if the null hypothesis is true and “far from” zero if the alternative hypothesis is true. The value of this test statistic for the creativity scores is 4.14. Is this “close to” or “far from” zero? The answer to that question comes from what is known about how the test statistic might have turned out in other randomization outcomes if there were no effect of the treatments.

Randomization Distribution of the Test Statistic

If the questionnaires had no effect on creativity, the subjects would have received exactly the same creativity scores regardless of the group to which they were assigned. Therefore, it is possible to determine what test statistic values would have occurred had the randomization process turned out differently. Display 1.7 has an example. The first column lists the creativity scores of all 47 participants in the study. The second column lists the groups to which they were assigned. The third column lists another possible way that the group assignments could have turned out. With that grouping, the difference in averages is 2.07.

It is conceptually possible to calculate $\bar{Y}_2 - \bar{Y}_1$ for every possible outcome of the randomization process (if there is no treatment effect). A histogram of all these values describes the *randomization distribution* of $\bar{Y}_2 - \bar{Y}_1$ if the null hypothesis is true. The number of possible outcomes of the random assignment is prohibitively large in this example—there are 1.6×10^{13} different groupings, which would take a computer capable of one million assignments per second half a year to complete.

Creativity score	Actual grouping	Another grouping	Creativity score	Actual grouping	Another grouping
12.0	Intrinsic(2)	1	5.0	Extrinsic(1)	2
12.0	Intrinsic	2	5.4	Extrinsic	2
12.9	Intrinsic	1	6.1	Extrinsic	1
13.6	Intrinsic	2	10.9	Extrinsic	2
16.6	Intrinsic	2	11.8	Extrinsic	1
17.2	Intrinsic	1	12.0	Extrinsic	1
17.5	Intrinsic	2	12.3	Extrinsic	1
18.2	Intrinsic	2	14.8	Extrinsic	2
19.1	Intrinsic	1	15.0	Extrinsic	2
19.3	Intrinsic	2	16.8	Extrinsic	2
19.8	Intrinsic	2	17.2	Extrinsic	2
20.3	Intrinsic	2	17.2	Extrinsic	1
20.5	Intrinsic	1	17.4	Extrinsic	2
20.6	Intrinsic	2	17.5	Extrinsic	2
21.3	Intrinsic	1	18.5	Extrinsic	2
21.6	Intrinsic	2	18.7	Extrinsic	1
22.1	Intrinsic	1	18.7	Extrinsic	1
22.2	Intrinsic	2	19.2	Extrinsic	1
22.6	Intrinsic	1	19.5	Extrinsic	1
23.1	Intrinsic	1	20.7	Extrinsic	1
24.0	Intrinsic	1	21.2	Extrinsic	1
24.3	Intrinsic	1	22.1	Extrinsic	2
26.7	Intrinsic	1	24.0	Extrinsic	2
29.7	Intrinsic	1			

↑
↑

Averages from actual grouping

Group	Average	Difference
Intrinsic (2)	19.88	
Extrinsic (1)	15.74	4.14

Averages from another grouping

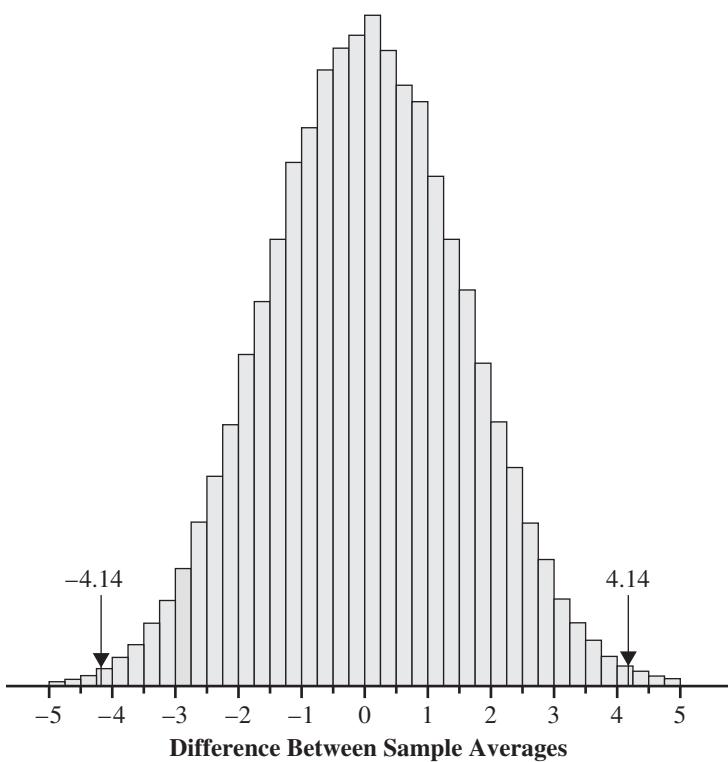
Group	Average	Difference
Group 1	18.87	
Group 2	16.80	2.07

Display 1.8, however, shows an estimate of the randomization distribution, obtained as the histogram of the values of $\bar{Y}_2 - \bar{Y}_1$ from 500,000 random regroupings of the numbers.

Display 1.8 suggests several things. First, it appears just as likely that test statistics will be negative as positive. Second, the majority of values fall in the range from -3.0 to $+3.0$. Third, only 1,335 of the 500,000 (0.27%) random regroupings produced test statistics as large as 4.14. This last point indicates that 4.14 is a value corresponding to an unusually uneven randomization outcome, if the null hypothesis is correct.

DISPLAY 1.8

Histogram of differences between group averages, from 500,000 regroupings of the creativity study data



The p-Value of the Test

Because the observed test statistic is improbably large under the null hypothesis, one is led to believe that the null hypothesis is wrong, that the treatment—not the randomization—caused the big difference. That conclusion could be incorrect, however, because the randomization is capable of producing such an extreme. The chance of its doing so is called the (*observed*) *p-value*. *In a randomized experiment the p-value is the probability that randomization alone leads to a test statistic as extreme as or more extreme than the one observed.* The smaller the *p*-value, the more unlikely it is that chance assignment is responsible for the discrepancy between groups, and the greater the evidence that the null hypothesis is incorrect. A general definition appears in Chapter 2.

Since 1,335 of the 500,000 regroupings produced differences as large or larger than the observed difference, the *p*-value for this study is estimated from the simulation to be $1,335/500,000 = 0.00267$. This is a *one-sided p-value* (for the alternative hypothesis that $\delta > 0$) because it counted as extremes only those outcomes with test statistics as large as or larger than the observed one. Statistics that are smaller than -4.14 may provide equally strong evidence against the null hypothesis, favoring the

alternative hypothesis that $\delta < 0$. If those (1,302 of them) are included, the result is a *two-sided* p -value of $2,637/500,000 = 0.005274$, which would be appropriate for the two-sided alternative hypothesis that $\delta \neq 0$.

Computing p -Values from Randomized Experiments

There are several methods for obtaining p -values in a randomized experiment. An enumeration of all possible regroupings of the data would represent all the ways that the data could turn out in all possible randomizations, absent any treatment effect. This would determine the answer exactly, but it is often overburdensome—as the creativity study illustrates. A second method, used previously, is to estimate the p -value by simulating a large number of randomizations and to find the proportion of these that produce a test statistic at least as extreme as the observed one.

The most common method is to approximate the randomization distribution with a mathematical curve, based on certain assumptions about the distribution of the measurements and the form of the test statistic. This approach is given formally in Chapter 2 along with a related discussion of confidence intervals for treatment effects.

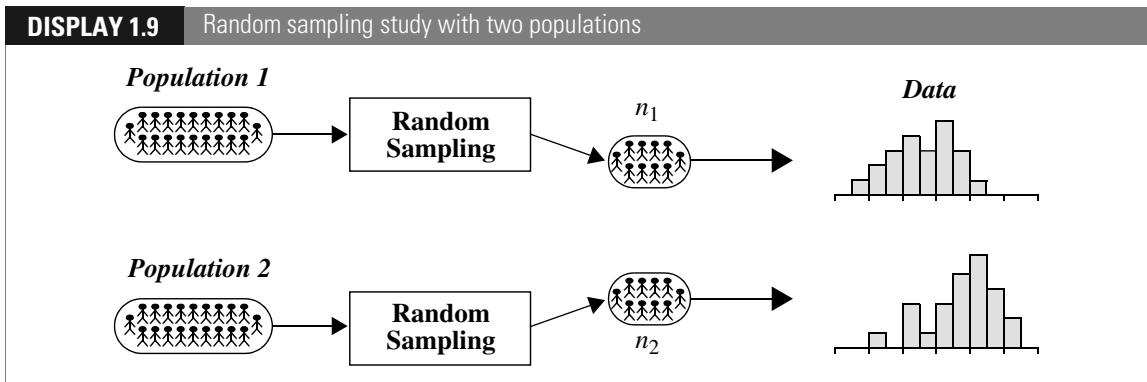
1.4 MEASURING UNCERTAINTY IN OBSERVATIONAL STUDIES

The p -value from the randomization test is a probability statement that is tied directly to the chance mechanism used to assign subjects to treatment groups. For observational studies there is no such chance mechanism for group assignment, so the thinking about statistical inference has to be different. Two approaches for observational studies are discussed here.

Section 1.4.1 introduces a probability model that connects the distribution of a test statistic to a chance mechanism for *sample selection* (which, remember, is something quite different from group assignment). In the sex discrimination study, however, the subjects were not selected randomly from any populations and, in fact, constitute the *entire* populations of males and females in their particular job category. Section 1.4.2 applies the same “regrouping” idea that led to the randomization test in order to say whether the salaries in the two groups differed more than would be expected from a random assignment of salaries to employees. Since this approach is tied to a fictitious model and not a bona fide probability model stemming from an actual chance mechanism, the inference is substantially weaker than the one from the randomized experiment. This is such an important distinction that the test goes by a different name—the permutation test—even though the calculations are identical to those of the randomization test.

1.4.1 A Probability Model for Random Sampling

Display 1.9 depicts random sampling from two populations. A chance mechanism, such as a lottery, selects a subset of n_1 units from population 1 in such a way that all subsets of size n_1 have the same chance of selection. The mechanism is used



again to select a random sample of size n_2 from population 2. The samples are drawn independently; that is, the particular sample drawn from one population does not influence the process of selection from the other.

Questions of interest in sampling studies center on how features of the populations differ. In many cases, the difference between the populations can be described by the difference between their means. If the means of populations 1 and 2 are μ_1 and μ_2 , respectively, then $\mu_2 - \mu_1$ becomes the single parameter for answering the questions of interest. Inferences and uncertainty measures are based on the difference, $\bar{Y}_2 - \bar{Y}_1$, between sample averages, which estimates the difference in population means.

As in Display 1.8, a *sampling distribution* for the statistic $\bar{Y}_2 - \bar{Y}_1$ is represented by a histogram of all values for the statistic from all possible samples that can be drawn from the two populations. The p -value for testing a hypothesis and the confidence intervals for estimating the parameter follow from an understanding of the sampling distribution. Full exposition appears in Chapter 2.

1.4.2 Testing for a Difference in the Sex Discrimination Study

In the sex discrimination study, there is no interest in the starting salaries of some larger population of individuals who were never hired, so a random sampling model is not relevant. Similarly, it makes no sense to view the sex of these individuals as randomly assigned. Neither the random sampling nor the randomized experiment model applies. Any interpretation that supports a statistical analysis must be based on a fictitious chance model.

One fictitious probability model for examining the difference between the average starting salary given to males and the average starting salary given to females assumes the employer assigned the set of starting salaries to the employees *at random*. That is, the employer shuffled a set of cards, each having one of the starting salaries written on it, and dealt them out. With this model, the investigator may ask whether the observed difference is a likely outcome.

Permutation Distribution of a Test Statistic

The collection of differences in averages from all possible assignments of starting salaries to individuals makes up the *permutation distribution* of the test statistic for this model. There are 8.7×10^{24} outcomes, so the million-a-second computer would now struggle along for about 275 million years. Shortcuts are available, however. It is not difficult to approximate the actual permutation distribution of a statistic by random regroupings of the data, as illustrated for the randomization distribution in the creativity study in Display 1.8. In addition, the easier-to-use *t-tools*, which will be introduced in Chapter 2, often provide very good approximations to the permutation distribution inferences. That approach was used to draw the statistical conclusion reported in Section 1.1.2. The *p*-value for comparing male and female distributions was less than 0.00001, indicating that the difference between male and female salaries was greater than what can reasonably be explained by chance. *It is calculated using the t-tools (see Chapter 2), which provide a close approximation.*

Scope of the Inferences

Inferences from fictitious models must be stated carefully. What the statistical argument shows is that the employer did not assign starting salaries at random—which was known at the outset. The strength of the argument comes from the demonstration that the actual assignment differs substantially from the expected outcome in *one model* where salary assignment is sex-blind.

1.5 RELATED ISSUES

1.5.1 Graphical Methods

This text relies heavily on the use of graphical methods for making decisions regarding the choice of statistical models. This section briefly reviews some important tools for displaying sets of numbers.

Relative Frequency Histograms

Histograms, like those in Display 1.4, are standard tools for displaying the general characteristics of a data set. Looking at the display, one should conclude that the central male starting salary is around \$6,000, that the central female starting salary is around \$5,200, that the two distributions have about the same spread—most salaries are within about \$1,000 of the centers—that both shapes appear reasonably symmetric, but that there is one male starting salary that appears very high in relation to the rest of the male salaries.

A histogram is a graph where the horizontal axis displays ranges for the measurement and the vertical axis displays the relative frequency per unit of measurement. Relative frequency is therefore depicted by area.

Other features are of less concern. Choices of the number of ranges (*bins*) and their boundaries can have some influence on the final picture. However, a

histogram is ordinarily used to show broad features, not exquisite detail, and the broad features will be apparent with many choices.

Stem-and-Leaf Diagrams

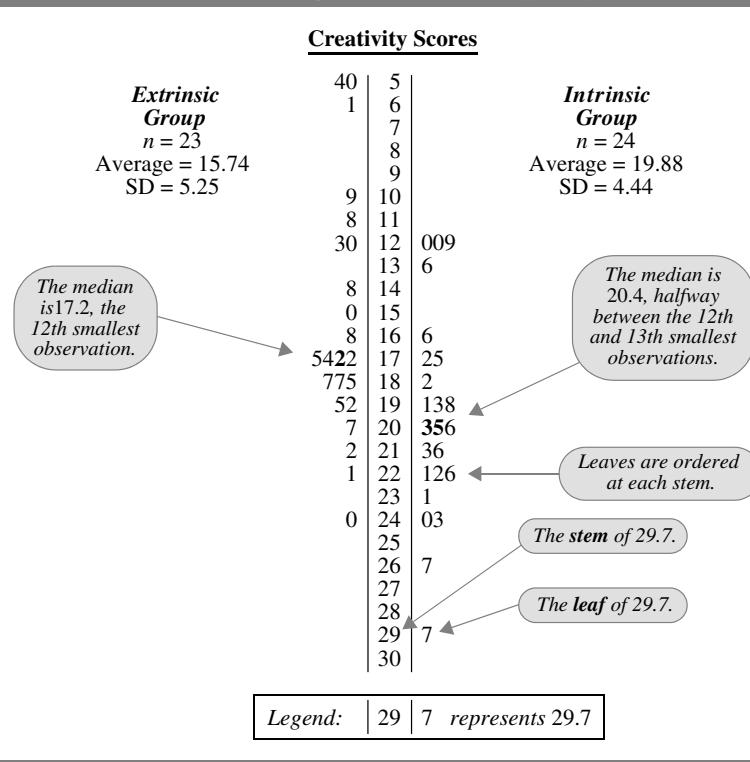
A *stem-and-leaf diagram* is a cross between a graph and a table. It is used to get a quick idea of the distribution of a set of measurements with pencil and paper or to present a set of numbers in a report.

Display 1.10 shows stem-and-leaf diagrams for the two sets of creativity scores. The digits in each observation are separated into a stem and a leaf. Each number in a set is represented in the diagram by its leaf on the same line as its stem. All possible stem values are listed in increasing order from top to bottom, whether or not there are observations with those stems. At each stem, all corresponding leaves are listed, in increasing order. *Outliers* may require a break in the string of stems.

The stem-and-leaf diagrams show the centers, spreads, and shapes of distributions in the same way histograms do. Their advantages include exact depiction of each number, ease of determining the median and quartiles of each set, and ease of construction. Disadvantages include difficulty in comparing distributions when the numbers of observations in data sets are very different and severe clutter with large data sets.

DISPLAY 1.10

Back-to-back stem-and-leaf diagrams for the creativity study data



Finding the Median from a Stem-and-Leaf Diagram

Because stem-and-leaf diagrams show the order of numbers in a set, they offer a convenient presentation for locating the *median* of the data. At least half the values in the data are greater than or equal to the median; at least half are less than or equal to it. A median is found by the “ $(n + 1)/2$ ” rule: calculate $k = (n + 1)/2$; if k is an integer, the median is the k th smallest observation; if k is halfway between two integers, the median is halfway between the corresponding observations (see Display 1.10).

Box Plots

A *box plot*, or box-and-whisker plot, is a graphical display that represents the middle 50% of a group of measurements by a box and highlights various features of the upper and lower 25% by other symbols. The box represents the body of the distribution of numbers, and the whiskers represent its tails. The graph gives an uncluttered view of the center, the spread, and the skewness of the distribution and indicates the presence of unusually small (or large) outlying values. Box plots are particularly useful for comparing several samples side by side. Unlike stem-and-leaf diagrams, box plots are not typically drawn by hand, but by a statistical computing program.

Most box plots use the *interquartile range* (IQR) to distinguish distant values from those close to the body of the distribution. The IQR is the difference between the upper and lower quartiles—the range of the middle 50% of the data.

Display 1.11 shows a box plot of gross domestic product (GDP) per capita in 228 countries. (See Exercise 16.) The lower quartile is \$2,875 per year and the upper quartile is \$22,250. The median income is marked by a line through the box at \$9,300 per year. Whiskers extend from the box out through all values that are within 1.5 IQRs of the box.

Observations that are more than 1.5 IQRs away from the box are far enough from the main body of data that they are indicated separately by dots. Observations more than 3 IQRs from the box are quite distant from the main body of data and are prominently marked, as in Display 1.11, and often named.

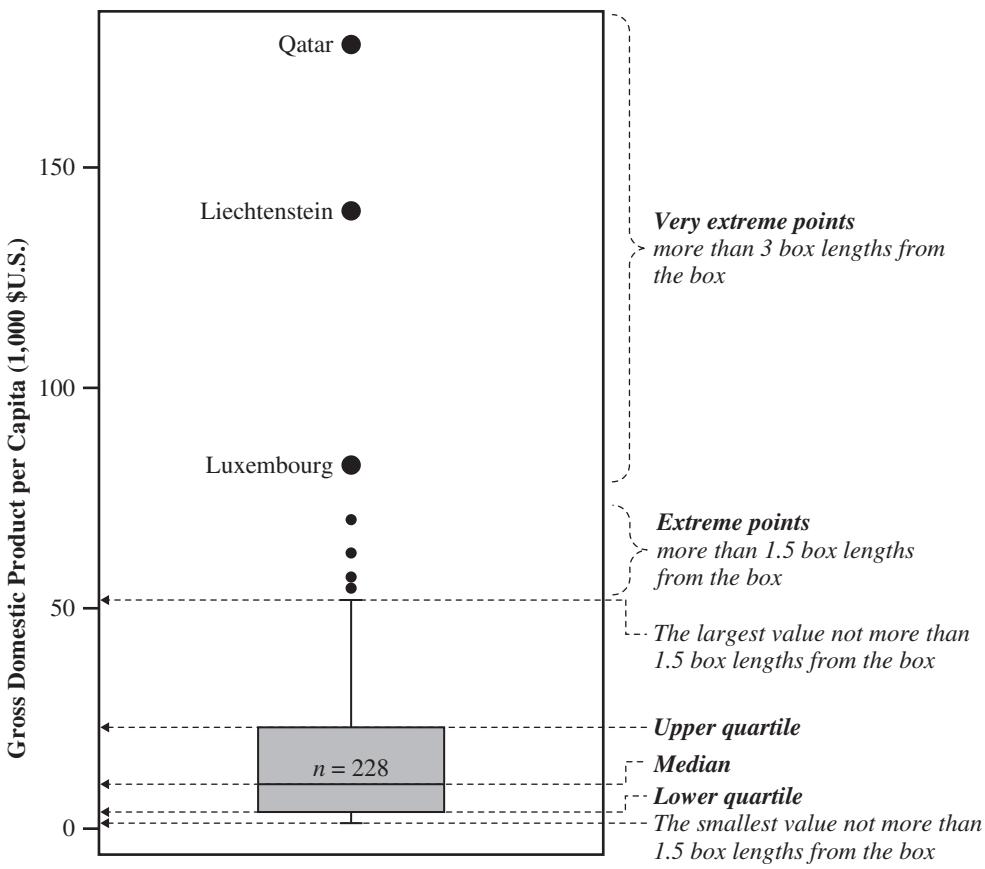
The difference in the distributions of starting salaries for men and women is shown by placing box plots *side by side*, as in Display 1.12. One can see at a glance that the middle starting salaries for men were about \$750 higher than for women, that the ranges of starting salaries were about the same for men and for women, and that the single starting salary of \$8,100 is unusually large.

Notes

1. Box-plotting routines are widely available in statistical computing packages. The definitions of quartiles, extreme points, and very extreme points, however, may differ slightly among packages.
2. The choices of 1.5 and 3 IQRs are arbitrary but useful cutoffs for highlighting observations distant and quite distant from the main body.
3. The width of the box is chosen simply with a view toward making the box look nice; it does not represent any aspect of the data.

DISPLAY 1.11

Box plot of per capita GDP for 228 countries in 2010 (\$U.S.)



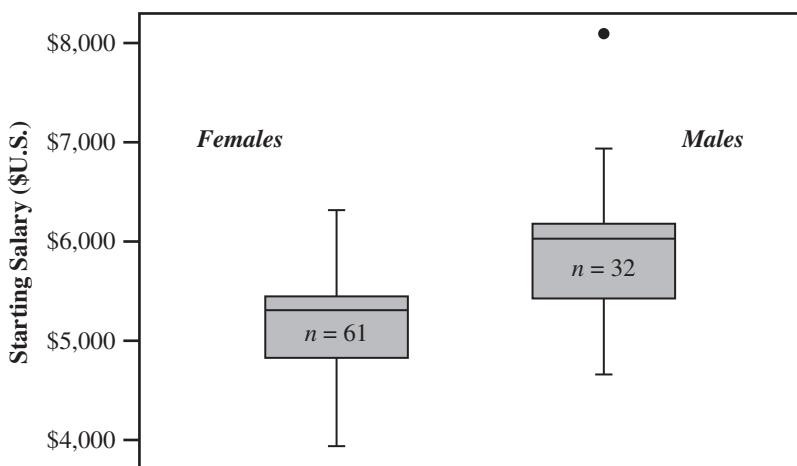
4. For presentation in a journal, it is common to define the box plot so that the whiskers extend to the 10th and 90th percentiles of the sample. This is easier to explain to readers who might be unfamiliar with box plots.
5. The data sets in Display 1.13 show a variety of different kinds of distributions. Box plots are matched with histograms of the same data sets.
6. Some statistical computer packages draw horizontal box plots.

1.5.2 Standard Statistical Terminology

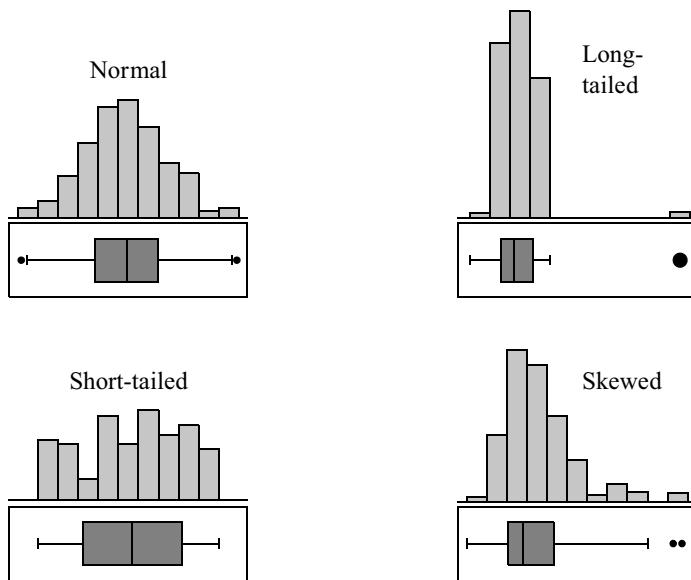
A *parameter* is an unknown numerical value describing a feature of a probability model. Parameters are indicated by Greek letters. A *statistic* is any quantity that can be calculated from the observed data. Statistics are represented by Roman letter symbols. An *estimate* is a statistic used as a guess for the value of a parameter. The notation for the estimate of a parameter θ is the parameter symbol with a hat on

DISPLAY 1.12

Side-by-side box plots for the starting salary data

**DISPLAY 1.13**

Histograms and box plots for 100 observations from four distributions



it, $\hat{\theta}$. Remember that the estimate can be calculated, but the parameter remains unknown.

The *Statistical Sleuth* uses the word *mean* when referring to an average calculated over an entire population. A mean is therefore a parameter. When referring to the average in a sample—which is both a statistic and an estimate of the population

mean—the *Statistical Sleuth* uses the word *average*. Standard notation uses μ for the mean and \bar{Y} for an average.

The *standard deviation* of a set of numbers Y_1, \dots, Y_n is defined as

$$\sqrt{\left(\sum_{i=1}^n (Y_i - \bar{Y})^2 \right) / (n - 1)}.$$

The symbol for a population standard deviation is σ ; for a sample standard deviation, the symbol is s . The standard deviation is a measure of spread, interpreted as the typical distance between a single number and the set's average.

1.5.3 Randomization of Experimental Units to Treatments

Experimental units are the human or animal subjects, plots of land, samples of material, or other entities that are assigned to treatment groups in an experiment and to which a treatment is applied. In some cases, the experimental units are groups of individuals, such as classrooms of students that are assigned, as a class, to receive one of several teaching methods; the classrooms—not the students—are the experimental units in this case. In the creativity example, the experimental units were the 47 people. The treatments applied to them were the motivation questionnaires.

In a two-treatment randomized experiment, all available experimental units have the same chance of being placed in group 1. Drawing cards from a thoroughly shuffled deck of red and black cards is one method. Tables of random numbers or computer-generated random numbers may also be used. These have the advantage that the randomization can be documented and checked against certain patterns.

Most statistical computer programs have routines that generate random numbers. To use such a list for randomly assigning units to two experimental groups, first list the units (in any order). Then generate a series of random numbers, assigning them down the list to the units. Reorder the list so that the random numbers are listed by increasing order, keeping each subject matched with its random number. Then assign the subjects with the smallest random numbers (at the top of the reordered list) to treatment 1 and those with the largest random numbers (at the bottom) to treatment 2. Ties in the random numbers can be broken by generating additional random numbers.

1.5.4 Selecting a Simple Random Sample from a Population

A *simple random sample* (of size n) is a subset of a population obtained by a procedure giving all sets of n distinct items in the population an equal chance of being chosen. To accomplish this in practice requires a *frame*: a numbered list of all the subjects. If a population of 20,000 members is so listed, an integer from 1 to 20,000 is associated with each. If a random sample of 100 is desired, it is only necessary to generate 100 distinct random integers such that each of the integers 1 through 20,000 has an equal chance of being selected. This can be accomplished with a statistical computer program.

Other Random Sampling Procedures

A commonly used method for selecting a sample of 100 from the list of 20,000 is to pick a single random start from the positions 1 to 200 and then to select every 200th subject going down the frame. This is called *systematic random sampling*. Another method is to group the frame into 200 blocks of 100 consecutive subjects each and to select one of the blocks at random. This is called *random cluster sampling*. In sampling units such as lakes of different sizes, it is sometimes useful to allow larger units to have higher probabilities of being sampled than smaller units. This constitutes *variable probability sampling*. These and other random sampling schemes can be useful, but they differ from simple random sampling in fundamental ways.

1.5.5 On Being Representative

Random samples and randomized experiments are representative in the same sense that flipping a coin to see who takes out the garbage is fair. Flipping is always fair before the coin hits the table, but the outcome is always unfair to the loser. In the same way, close examination of the results of randomization or random sampling can usually expose ways in which the chosen sample is not representative. The key, however, is not to abandon the procedure when its result is suspect. Uncertainty about representativeness will be incorporated into the statistical analysis itself. If randomization were abandoned, there would be no way to express uncertainty accurately.

1.6 SUMMARY

Cause-and-effect relationships can be inferred from randomized experiments but not from observational studies. The problem with observational studies is that confounding variables—identifiable or not—may be responsible for observed differences. Randomized experiments eliminate this problem by ensuring that differences between groups (other than those of the assigned treatments) are due to chance alone. Statistical measures of uncertainty account for this chance.

Statistically, the statements that generalize sample results to more general contexts are based on a probability model. When the model corresponds to the planned use of randomization or random sampling, it provides a firm basis for drawing inferences. The probability model may also be a fiction, created for the purposes of assessing uncertainty. Results from fictitious probability models must be viewed with skepticism.

1.7 EXERCISES

Conceptual Exercises

- 1. Creativity Study.** In the motivation and creativity experiment, the poems were given to the judges in random order. Why was that important?

- 2. Sex Discrimination Study.** Explain why it is difficult to prove sex discrimination (that males in a company receive higher starting salaries because they are males) even if it has occurred.
- 3.** A study found that individuals who lived in houses with more than two bathrooms tended to have higher blood pressure than individuals who lived in houses with two or fewer bathrooms. (a) Can cause and effect be inferred from this? (b) What confounding variables may be responsible for the difference?
- 4.** A researcher performed a comparative experiment on laboratory rats. Rats were assigned to group 1 haphazardly by pulling them out of the cage without thinking about which one to select. Should others question the claim that this was as good as a randomized experiment?
- 5.** In 1930 an experiment was conducted on 20,000 school children in England. Teachers were responsible for randomly assigning their students to a treatment group—to receive $\frac{3}{4}$ pint of milk each day—or to a control group—to receive no milk supplement. Weights and heights were measured before and after the four-month experiment. The study found that children receiving milk gained more weight during the study period. On further investigation, it was also found that the controls were heavier and taller than the treatment group *before* the milk treatment began (more so than could be attributed to chance). What is the likely explanation and the implication concerning the validity of the experiment?
- 6.** Ten marijuana users, aged 14 to 16, were drawn from patients enrolled in a drug abuse program and compared to nine drug-free volunteers of the same age group. Neuropsychological tests for short-term memory were given, and the marijuana group average was found to be significantly lower than the control group average. The marijuana group was held drug-free for the next six weeks, at which time a similar test was given with essentially the same result. The researchers concluded that marijuana use caused adolescents to have short-term memory deficits that continue for at least six weeks after the last use of marijuana. (a) Can a genuine causal relationship be established from this study? (b) Can the results be generalized to other 14- to 16-year-olds? (c) What are some potential confounding factors?
- 7.** Suppose that random samples of Caucasian-American and Chinese-American individuals are obtained from patient records of doctors participating in a study to compare blood pressures of the two populations. Suppose that the individuals selected are asked whether they want to participate and that some decline. The study is conducted only on those that volunteer to participate, and a comparison of the distributions of blood pressures is conducted. Where does this study fit in Display 1.5? What assumption would be necessary to allow inferences to be made to the sampled populations?
- 8.** More people get colds during cold weather than during warm weather. Does that prove that cold temperatures cause people to get colds? What is a potential confounding factor?
- 9.** A study showed that children who watch more than two hours of television each day tend to have higher cholesterol levels than children who watch less than two hours of television each day. Can you think of any use for the result of this study?
- 10.** What is the difference between a randomized experiment and a random sample?
- 11.** A number of volunteers were randomly assigned to one of two groups, one of which received daily doses of vitamin C and one of which received daily placebos (without any active ingredient). It was found that the rate of colds was lower in the vitamin C group than in the placebo group. It became evident, however, that many of the subjects in the vitamin C group correctly guessed that they were receiving vitamin C rather than placebo, because of the taste. Can it still be said that the difference in treatments is what caused the difference in cold rates?
- 12. Fish Oil and Blood Pressure.** Researchers used 7 red and 7 black playing cards to randomly assign 14 volunteer males with high blood pressure to one of two diets for four weeks: a fish oil diet and a standard oil diet. The reductions in diastolic blood pressure are shown in Display 1.14.

DISPLAY 1.14

Reductions in diastolic blood pressure (mm of mercury) for 14 men after 4 weeks of a special diet containing fish oil or a regular, nonfish oil

Fish oil diet:	8	12	10	14	2	0	0
Regular oil diet:	-6	0	1	2	-3	-4	2

(Based on a study by H. R. Knapp and G. A. FitzGerald, “The Antihypertensive Effects of Fish Oil,” *New England Journal of Medicine* 320 (1989): 1037–43.) Why might the results of this study be important, even though the volunteers do not constitute a random sample from any population?

13. Why does a stem-and-leaf diagram require less space than an ordinary table?
14. What governs the *width* of a box plot?
15. What general features are evident in a box plot of data from a normal distribution? from a skewed distribution? from a short-tailed distribution? from a long-tailed distribution?

Computational Exercises

16. **Gross Domestic Product (GDP) per Capita.** The data file ex0116 contains the gross domestic product per capita for 228 countries—the data used to construct the box plot in Display 1.11. (a) Make a box plot of the per capita GDPs with a statistical computer program. Include a *y*-axis label (for example, “Gross Domestic Product per Capita in U.S.”). (b) In what ways, if any, is the display from your software different from Display 1.11? (c) Use a statistical computer program to draw a *histogram* of the per capita GDPs. Include an *x*-axis label. Use the program’s default bin width. Report that bin width. (d) The program’s default bin width for histograms is not necessarily the best choice. If it’s possible with your statistical computer program, redraw the histogram of part (c) using bin widths of \$5,000. (Data from Central Intelligence Agency, “Country Comparison: GDP—per capita (PPP),” *The World Factbook*, June 24, 2011 <https://www.cia.gov/library/publications/the-world-factbook/rankorder/2004rank.html> (June 30, 2011).)
17. Seven students volunteered for a comparison of study guides for an advanced course in mathematics. They were randomly assigned, four to study guide A and three to study guide B. All were instructed to study independently. Following a two-day study period, all students were given an examination about the material covered by the guides, with the following results:

Study Guide A scores: 68, 77, 82, 85

Study Guide B scores: 53, 64, 71

Perform a randomization test by listing all possible ways that these students could have been randomized to two groups. There are 35 ways. For each outcome, calculate the difference between sample averages. Finally, calculate the two-sided *p*-value for the observed outcome.

18. Using the creativity study data (Section 1.1.1) and a computer, assign a set of random numbers to the 47 subjects. Order the entire data set by increasing values of the random numbers, and divide the subjects into group 1 with the 24 lowest random numbers and group 2 with the 23 highest. Compute the difference in averages. Repeat this process five times, using different sets of random numbers. Did you get any differences larger than the one actually observed (4.14)?
19. Write down the names and ages of 10 people. Using coin flips, divide them into two groups, as if for a randomized experiment. Did one group tend to get many of the older subjects? Was there any way to predict which group would have a higher average age in advance of the coin flips?

- 20.** Repeat Exercise 19 using a randomization mechanism that ensures that each group will end up with exactly five people.
- 21.** Read the methods and design sections of five published studies in your own field of specialization. (a) Categorize each according to Display 1.5. (b) Now read the conclusions sections. Determine whether inferential statements are limited to or go beyond the scope allowed in Display 1.5.
- 22.** Draw back-to-back stem-and-leaf diagrams of the creativity scores for the two motivation groups in Display 1.1 (by hand).
- 23.** Use the computer to draw side-by-side box plots for the creativity scores in Display 1.1.
- 24.** Using the stem-and-leaf diagrams from Exercise 22, compute the median, lower quartile, and upper quartile for each of the motivation groups. Identify these on the box plots from Exercise 23. Using a ruler, measure the length of the box (the IQR) for each group, and make horizontal lines at 1.5 IQRs above and below each box and at 3 IQRs above and below the box. Are there any *extreme points* in either group? Are there any *very extreme points* in either group?
- 25.** The following are zinc concentrations (in mg/ml) in the blood for two groups of rats. Group A received a dietary supplement of calcium, and group B did not. Researchers are interested in variations in zinc level as a side effect of dietary supplementation of calcium.

Group A: 1.31 1.45 1.12 1.16 1.30 1.50 1.20 1.22 1.42 1.14 1.23 1.59 1.11 1.10
1.53 1.52 1.17 1.49 1.62 1.29

Group B: 1.13 1.71 1.39 1.15 1.33 1.00 1.03 1.68 1.76 1.55 1.34 1.47 1.74 1.74
1.19 1.15 1.20 1.59 1.47

- (a) Make a back-to-back stem-and-leaf diagram (by hand) for these measurements. (b) Use the computer to draw side-by-side box plots.

Data Problems

- 26. Environmental Voting of Democrats and Republicans in the U.S. House of Representatives.** Each year, the League of Conservation Voters (LCV) identifies legislative votes taken in each house of the U.S. Congress—votes that are highly influential in establishing policy and action on environmental problems. The LCV then publishes whether each member of Congress cast a pro-environment or an anti-environment vote. Display 1.15 shows these votes during the years 2005, 2006, and 2007 for members of the House of Representatives. Evaluate the evidence supporting party differences in the percentage of pro-environment votes. Write a brief report of your conclusion, including a graphical display and summary statistics.

DISPLAY 1.15

Number of pro- and anti-environment votes in 2005, 2006, and 2007, according to the League of Conservation Voters, of Republican (R) and Democratic (D) members of the U.S. House of Representatives; and the total percentage of their votes that were deemed pro-environment; first 5 of 492 rows

State	Representative	Party	Pro05	Anti05	Pro06	Anti06	Pro07	Anti07	PctPro
Alabama	Bonner	R	2	16	3	9	2	18	14.0
Alabama	Everett	R	0	18	1	11	2	18	6.0
Alabama	Rogers	R	1	17	2	10	3	17	12.0
Alabama	Aderholt	R	0	18	0	12	2	18	4.0
Alabama	Cramer	D	5	13	4	7	14	6	46.9

27. Environmental Voting of Democrats and Republicans in the U.S. Senate. Display 1.16 shows the first five rows of a data set with pro- and anti-environment votes (according to the League of Conservation Voters; see Exercise 26) during 2005, 2006, and 2007, cast by U.S. senators. Evaluate the evidence supporting party differences in the percentage of pro-environment votes. Write a brief report of your conclusion, and include a graphical display and summary statistics.

DISPLAY 1.16

Number of pro- and anti-environment votes in 2005, 2006, and 2007, according to the League of Conservation Voters, of Republican (R) and Democratic (D) members of the U.S. Senate; and the total percentage of their votes that were deemed pro-environment; first 5 of 112 rows

State	Senator	Party	Pro05	Anti05	Pro06	Anti06	Pro07	Anti07	PctPro
Alabama	Session	R	1	18	0	7	2	13	7.3
Alabama	Shelby	R	1	19	0	7	1	14	4.8
Alaska	Murkowski	R	2	17	1	6	6	9	22.0
Alaska	Stevens	R	1	19	1	6	4	11	14.3
Arizona	Kyle	R	1	19	2	5	2	13	11.9

Answers to Conceptual Exercises

- If one of the two treatment groups had their poems judged first, then the effect of motivation treatment would be confounded with time effects in the judges' marking of creativity scores. Judges are influenced by memories of previous cases. They may also discern a change in average quality, which could alter their expectations.
- Statistically, the distributions of male and female starting salaries may be compared after adjusting for possible confounding variables. Since the data are necessarily observational, a difference in distribution cannot be linked to a specific cause. Once more, for emphasis: The best possible statistical analysis using the best possible data cannot establish causation in an observational study, but observational data are the only data likely to be available for discrimination cases. If, therefore, courts required plaintiffs to produce scientifically defensible proof of discrimination in order to prevail, defendants would win all discrimination cases *by definition*. As a result, some courts that wish to give weight to statistical information in discrimination cases adopt rules of evidence that allow proof to be established negatively—by the lack of an adequate rebuttal.
- (a) No. (b) Wealth and richness of diet.
- Yes. First of all, there is the possibility that the rats that were easier to get out of the cage are different from the others—bigger, less mobile, perhaps. Second, even if there is no obvious reason why the rats in the two groups might be different, that does not ensure they are not. Researchers must maintain skepticism about the conclusions in light of the uncertainty. With proper randomization, which is easy to carry out, there would be no doubt.
- Teachers apparently gave the milk to the students they thought would most benefit from it. As a consequence, the results of the experiment were not valid.
- (a) No. In this observational study, it is possible that the drug users were different from the nonusers in ways other than drug use. (b) No, because the samples are volunteers, not random samples. (c) Happiness of the child, stability of family, success of child in school.

7. It is an observational study. The population that is randomly sampled is the population of consenting subjects. To draw inference to all subjects requires the assumption that the response is unrelated to the reasons for consent.
8. No. The amount of time people spend indoors.
9. It may be possible for doctors to identify children who are likely to have high cholesterol by asking about their television watching habits. This requires no causative link. It is a minor use, however, and there is apparently no other.
10. In a randomized experiment a random mechanism is used to allocate the available subjects to treatment groups. In a random sample a random mechanism is used to select subjects from the populations of interest.
11. Yes, sort of. The treatment difference caused the different responses, but the actual “treatment” received in the vitamin C group was both a daily dose of vitamin C and knowledge that it was vitamin C. It’s possible that the second aspect of this treatment is what was responsible for the difference. Researchers must make sure that the two groups are treated as similarly as possible in all respects, except for the specific agent under comparison.
12. The conclusion that the fish oil diet causes a reduction in blood pressure for these volunteers is a strong and useful one, even if it formally applies only to these particular individuals.
13. The leading digit or digits (the stems) are listed only once.
14. The width of the box is chosen to make the overall picture pleasing to the eye. It does not represent anything. For side-by-side box plots, the widths of the two boxes should be equal.
15. Normal: Median line in the middle of the box; equal whiskers; few if any extreme points; no very extreme points. Skewed: extreme points in one direction only; very extreme points possible, but few; long whisker on side of extreme points and short whisker (if any) on other side; median line closer to short whisker than to long one. Short-tailed: Like normal with no extreme points and very short whiskers. Long-tailed: Like normal (roughly symmetric) with extreme points strung out in both tails; some very extreme points possible.

Inference Using *t*-Distributions

The first of the case studies in this chapter illustrates the structure of two independent samples and the second illustrates the structure of a single sample of pairs. This chapter has the dual purposes of detailing the inferential tools for these important structures and emphasizing the conceptual basis of confidence intervals and *p*-values based on the *t*-distribution more generally. These “*t*-tools” are useful in regression and analysis of variance structures that will be taken up in subsequent chapters, but their main features can be conveyed in the relatively simple setting of a two-group comparison. The tools are derived under random sampling models when populations are normally distributed. As will be seen, the resulting tools also find application as approximations to randomization tests. Furthermore, as described in Chapter 3, they often work quite well even when the populations are not normal.

2.1 CASE STUDIES

2.1.1 Evidence Supporting Darwin's Theory of Natural Selection—An Observational Study

In the search for evidence supporting Charles Darwin's theory of natural selection, Karl Pearson (1857–1936) made a simple yet profound contribution to human reasoning. Pearson was encouraged by Darwin's cousin, Francis Galton, to use mathematical analysis of numerical characteristics of animals to find evidence of species adaptation. Darwin had observed, for example, that beaks of various finch species in the Galápagos Islands differed in ways that seemed well suited to survival in their respective environments, but he had no direct evidence that the beak characteristics actually evolved to adapt to their environments. In thinking about the use of mathematics, Pearson puzzled over how to show that a numerical characteristic, such as beak size, differs in populations before and after an environmental disturbance if, in fact, the characteristic is *variable* among individuals within each population. His profound contribution was to articulate that the scientific questions could be addressed through a comparison of *population distributions* of the characteristic. Even if two populations of finches showed considerable overlap in beak sizes, he argued, a demonstration that the population distributions differed would suffice as evidence of a species difference. Although Pearson never found the evidence he was looking for, his insights and investigations into methodology for comparing population distributions had a major impact on the development of modern statistics.

In the 1980s, biologists Peter and Rosemary Grant and colleagues found what Pearson had been looking for. Over the course of 30 years, the Grants' research team caught and measured all the birds from more than 20 generations of finches on the Galápagos island of Daphne Major. In one of those years, 1977, a severe drought caused vegetation to wither, and the only remaining food source was a large, tough seed, which the finches ordinarily ignored. Were the birds with larger and stronger beaks for opening these tough seeds more likely to survive that year and did they tend to pass this characteristic to their offspring?

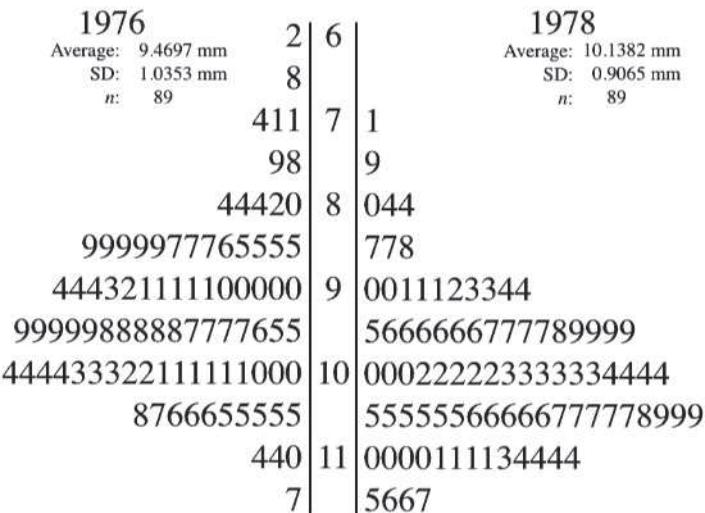
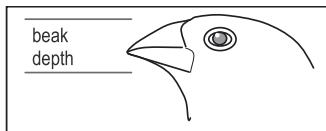
The Grants measured beak depths (height of the beak at its base) of all 751 Daphne Major finches the year before the drought (1976) and all 89 finches captured the year after the drought (1978). Display 2.1 shows side-by-side stem-and-leaf diagrams comparing the 89 post-drought finch bill depths with an equal-sized random sample of the pre-drought bill depths. (For the full set of 1976 finches, see Exercise 2.18.) Is there evidence of a difference between the population distributions of beak depths in 1976 and 1978? (The data were read from a histogram in P. Grant, 1986, *Ecology and Evolution of Darwin's Finches*, Princeton University Press, Princeton, N.J.)

Statistical Conclusion

These data provide overwhelming evidence that the mean beak depth increased from 1976 to 1978 (one-sided p -value < 0.00001 from a two-sample t -test). The

DISPLAY 2.1

Beak depths (mm) of Darwin finches on Daphne Major in 1976, pre-drought, and 1978, post-drought



Legend: | 11 | 0 = 11.0 mm

1978 (post-drought) mean was estimated to exceed the 1976 (pre-drought) mean by 0.67 mm (95% confidence interval: 0.38 mm to 0.96 mm).

Scope of Inference

Since this was an observational study, a causal conclusion—that the drought *caused* a change in the mean beak size—does not follow directly from the statistical evidence of a difference in the means. A lack of alternative explanations, though, might make biologists reasonably confident in the speculation that natural selection in response to the drought is responsible for the change. The Grants measured every finch in the 1976 and 1978 populations. Even though the entire populations were measured, a two-sample *t*-test is appropriate for showing that the difference in the two populations is greater than can be explained by chance. A more serious problem, though, is that the population in 1978 likely includes birds that were also in the 1976 population or offspring of birds in the 1976 population. If so, the assumption of independence of the two samples would be violated.

2.1.2 Anatomical Abnormalities Associated with Schizophrenia—An Observational Study

Are any physiological indicators associated with schizophrenia? Early studies, based largely on postmortem analysis, suggest that the sizes of certain areas of the brain may be different in persons afflicted with schizophrenia than in others. Confounding variables in these studies, however, clouded the issue considerably. In a 1990 article, researchers reported the results of a study that controlled for genetic and socioeconomic differences by examining 15 pairs of monozygotic twins, where one of the twins was schizophrenic and the other was not. The twins were located through an intensive search throughout Canada and the United States. (Data from R. L. Suddath et al., “Anatomical Abnormalities in the Brains of Monozygotic Twins Discordant for Schizophrenia,” *New England Journal of Medicine* 322(12) (1990): 789–93.)

The researchers used magnetic resonance imaging to measure the volumes (in cm^3) of several regions and subregions inside the twins’ brains. Display 2.2 presents data based on the reported summary statistics from one subregion, the left hippocampus. What is the magnitude of the difference in volumes of the left hippocampus between the unaffected and the affected individuals? Can the observed difference be attributed to chance?

Summary of Statistical Analysis

There is substantial evidence that the mean difference in left hippocampus volumes between schizophrenic individuals and their nonschizophrenic twins is nonzero (two-sided p -value = 0.006, from a paired t -test). It is estimated that the mean

DISPLAY 2.2

Differences in volumes (cm^3) of left hippocampus in 15 sets of monozygotic twins where one twin is affected by schizophrenia

Pair #	Unaffected	Affected	Difference	Differences		
1	1.94	1.27	0.67			Average: 0.199
2	1.44	1.63	-0.19	-2	9	Sample SD: 0.238
3	1.56	1.47	0.09	-1		<i>n</i> : 15
4	1.58	1.39	0.19	-0		
5	2.06	1.93	0.13	0	23479	
6	1.66	1.26	0.40	1	0139	
7	1.75	1.71	0.04	2	3	
8	1.77	1.67	0.10	3		
9	1.78	1.28	0.50	4	0	
10	1.92	1.85	0.07	5	09	
11	1.25	1.02	0.23	6	7	
12	1.93	1.34	0.59	7		
13	2.04	2.02	0.02			
14	1.62	1.59	0.03			
15	2.08	1.97	0.11			

Legend: | 6 | 7 represents 0.67 cm^3

volume is 0.20 cm^3 smaller for those with schizophrenia (about 11% smaller). A 95% confidence interval for the difference is from 0.07 to 0.33 cm^3 .

Scope of Inference

These twins were not randomly selected from general populations of schizophrenic and nonschizophrenic individuals. Tempting as it is to draw inferences to these wider populations, such inferences must be based on an assumption that these individuals are as representative as random samples are. Furthermore, the study is observational, so no causal connection between left hippocampus volume and schizophrenia can be established from the statistics alone. In fact, the researchers had no theories about whether the abnormalities preceded the disease or resulted from it.

2.2 ONE-SAMPLE *t*-TOOLS AND THE PAIRED *t*-TEST

The schizophrenia study used a *paired t*-test, in which measurements taken on paired subjects are reduced to a single set of differences for analysis. This section develops the single population methods for drawing inferences about the population mean from a random sample and at the same time introduces key concepts such as sampling distributions, standard errors, *Z*-ratios, and *t*-ratios.

2.2.1 The Sampling Distribution of a Sample Average

A random sample is drawn from a population with the objective of learning about the population's mean. Suppose the average of that sample is written on a piece of paper, which is placed in a box. Then suppose this process is repeated for every one of the equally likely samples that could be drawn. Then the distribution of all the numbers in the box is the *sampling distribution of the average*.

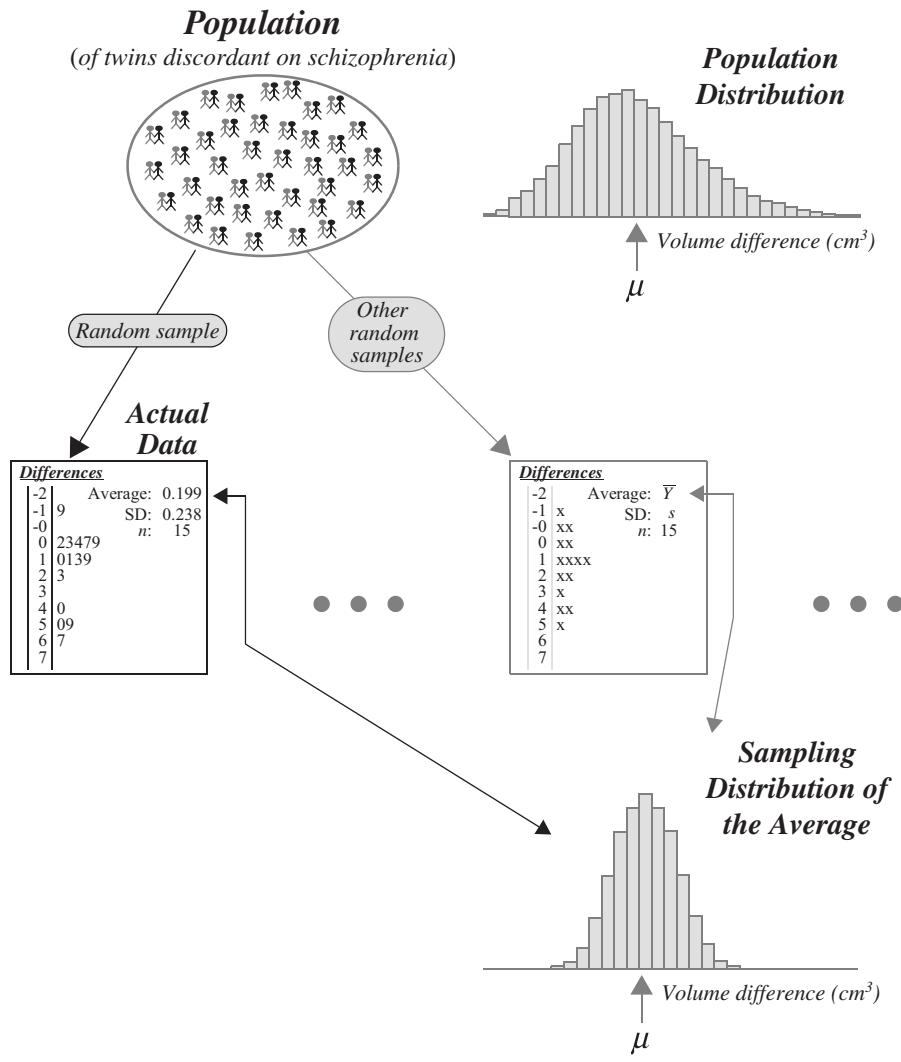
Display 2.3 illustrates a sampling distribution in the conceptual framework of the schizophrenia study. There is an assumed population of twins in which one of the twins has schizophrenia and the other does not. For each set of twins, Y represents the difference between the left hippocampus volumes of the unaffected and the affected twin. The 15 observed differences are assumed to be a random sample from this population. To examine whether there is a structural difference between volumes, one calculates the average of the 15 measurements, $\bar{Y} = 0.20\text{ cm}^3$, as an estimate of the population mean μ .

Although only the one sample is actually taken, it is important to think about replicating the study—repeatedly collecting 15 sets of twins and repeatedly calculating the average difference. The value of the average varies from sample to sample, and a histogram of its values represents its sampling distribution.

Because, in this case, there is only one average with which to estimate a population mean, it would seem difficult to learn anything about the characteristics of the sampling distribution. Some illuminating facts about the sampling distribution of an average, however, come from statistical theory. If a population has

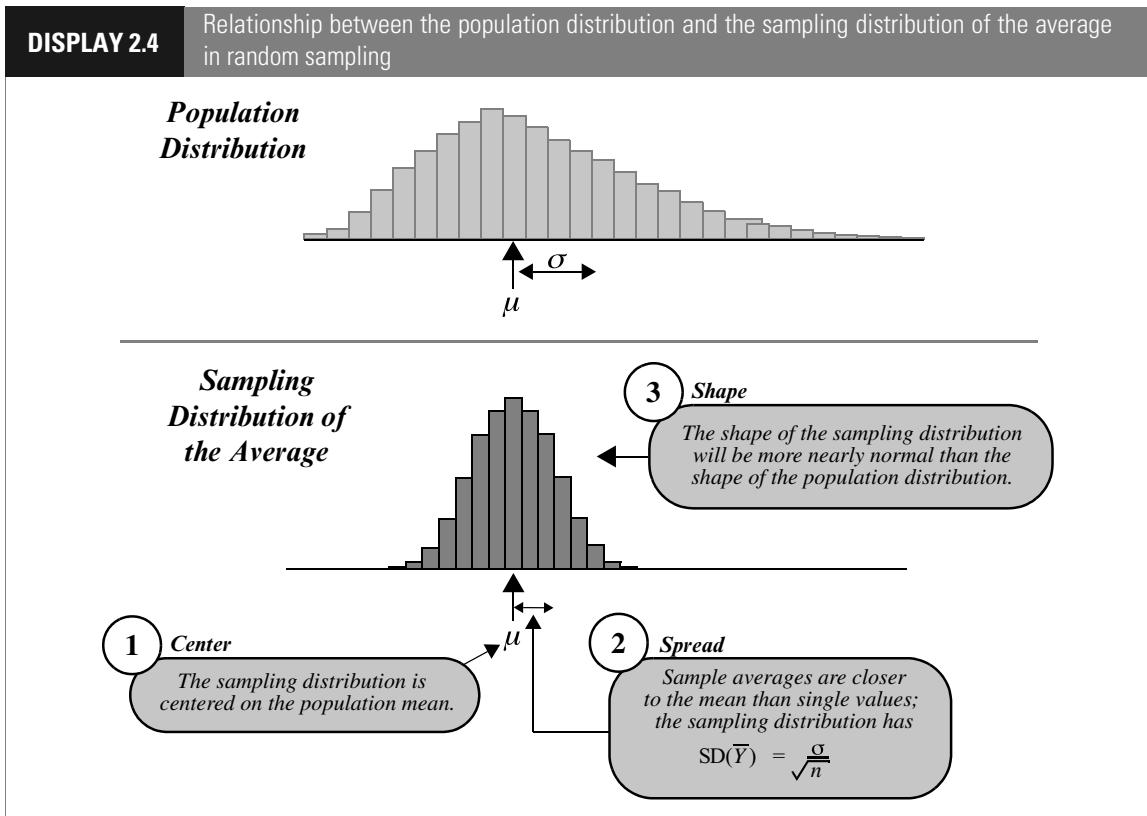
DISPLAY 2.3

The sampling distribution of the sample average



mean μ and standard deviation σ , then—as illustrated in Display 2.4—the mean of the sampling distribution of the average is also μ , the standard deviation of the sampling distribution is σ/\sqrt{n} , and the shape of the sampling distribution is more nearly normal than is the shape of the population distribution. The last fact comes from the important *Central Limit Theorem*.

The standard deviation in the sampling distribution of an average, denoted by $SD(\bar{Y})$, is the typical size of $(\bar{Y} - \mu)$, the error in using \bar{Y} as an estimate of μ . This standard deviation gets smaller as the sample size increases.



2.2.2 The Standard Error of an Average in Random Sampling

The *standard error* of any statistic is an estimate of the standard deviation in its sampling distribution. It is therefore the best guess about the likely size of the difference between a statistic used to estimate a parameter and the parameter itself. Standard errors are ordinarily calculated by substituting estimates of variability parameters into the formulas of sampling distribution standard deviations.

Associated with every standard error is a measure of the amount of information used to estimate variability, called its *degrees of freedom*, denoted d.f. Degrees of freedom are measured in units of “equivalent numbers of independent observations.” Further information to explain degrees of freedom, more generally, will be provided in future chapters.

Standard Error for a Sample Average

The formula for the standard deviation of the average in a sample of size n is σ/\sqrt{n} , so if s is the sample standard deviation, the standard error for the average is

$$\text{SE}(\bar{Y}) = \frac{s}{\sqrt{n}}, \quad \text{d.f.} = (n - 1).$$

The degrees of freedom in a single sample standard deviation are always one less than the sample size.

In the schizophrenia study, the average difference between volumes of unaffected and affected twins is 0.199 cm^3 , and the sample standard deviation of the differences is 0.238 cm^3 . The standard error of the average is therefore 0.062 cm^3 , with 14 degrees of freedom. From this, one makes a preliminary judgment that the population difference is likely to be near the sample estimate, 0.199 cm^3 , but that the sample estimate is likely to be somewhere in the neighborhood of 0.062 cm^3 off the mark.

2.2.3 The *t*-Ratio Based on a Sample Average

The ratio of an estimate's error to the anticipated size of its error provides a convenient basis for drawing inferences about the parameter in question.

The Z-Ratio

For any parameter and its sample estimate, its *Z-ratio* is defined as $Z = (\text{Estimate} - \text{Parameter})/\text{SD}(\text{Estimate})$. If the sampling distribution of the estimate is normal, then the sampling distribution of Z is *standard normal*, where the mean is 0 and the standard deviation is 1. The known percentiles of the standard normal distribution permit an understanding of the likely values of the *Z*-ratio, even though its value in any single case will not be known. From a computer program with normal percentiles, for example, it can be found that for 95% of all samples the *Z*-ratio will fall in the interval -1.96 to 1.96 . If the standard deviation of the estimate is known, this permits an understanding of the likely size of the estimation error. Consequently, useful statements can be made about the amount of uncertainty with which questions about the parameter can be resolved.

The t-Ratio

When, as is usually the case, the standard deviation of an estimate is unknown, it is natural to replace its value in the *Z*-ratio with the estimate's standard error. The result is the *t-ratio*,

$$t\text{-ratio} = \frac{(\text{Estimate} - \text{Parameter})}{\text{SE}(\text{Estimate})}$$

Associated with this *t*-ratio are the same degrees of freedom associated with the standard error of the estimate. The *t*-ratio does not have a standard normal

distribution, because there is extra variability due to estimating the standard deviation. The fewer the degrees of freedom, the greater is this extra variability. Under some conditions, however, the sampling distribution of the *t*-ratio is known.

If \bar{Y} is the average in a random sample of size n from a normally distributed population, the sampling distribution of its t-ratio is described by a Student's t-distribution on $n - 1$ degrees of freedom. The mathematical formula for the *t*-distributions was guessed by W. S. Gossett, a scientist who worked at the Guinness Brewery in the early 1900s and who published under the pseudonym "Student." The formula was proved to be correct by R. A. Fisher in 1925.

Histograms for *t*-distributions are symmetric about zero. For large degrees of freedom (about 30 or more), *t*-distributions differ very little from the standard normal. For smaller degrees of freedom, they have longer tails than normal. Percentiles of *t*-distributions are available in statistical computer programs and some pocket calculators.

2.2.4 Unraveling the *t*-Ratio

The average difference between hippocampus volumes for the 15 sets of twins in the schizophrenia study is 0.199 cm^3 , and its standard error is 0.0615 cm^3 , based on 14 degrees of freedom. The *t*-ratio is therefore $(0.199 - \mu)/0.0615$, where μ is the mean difference in the population of twins. If it can be assumed that the population distribution is normally distributed, then this *t*-ratio has a value typical of one drawn at random from a *t*-distribution with 14 degrees of freedom. A picture of this distribution is shown in Display 2.5.

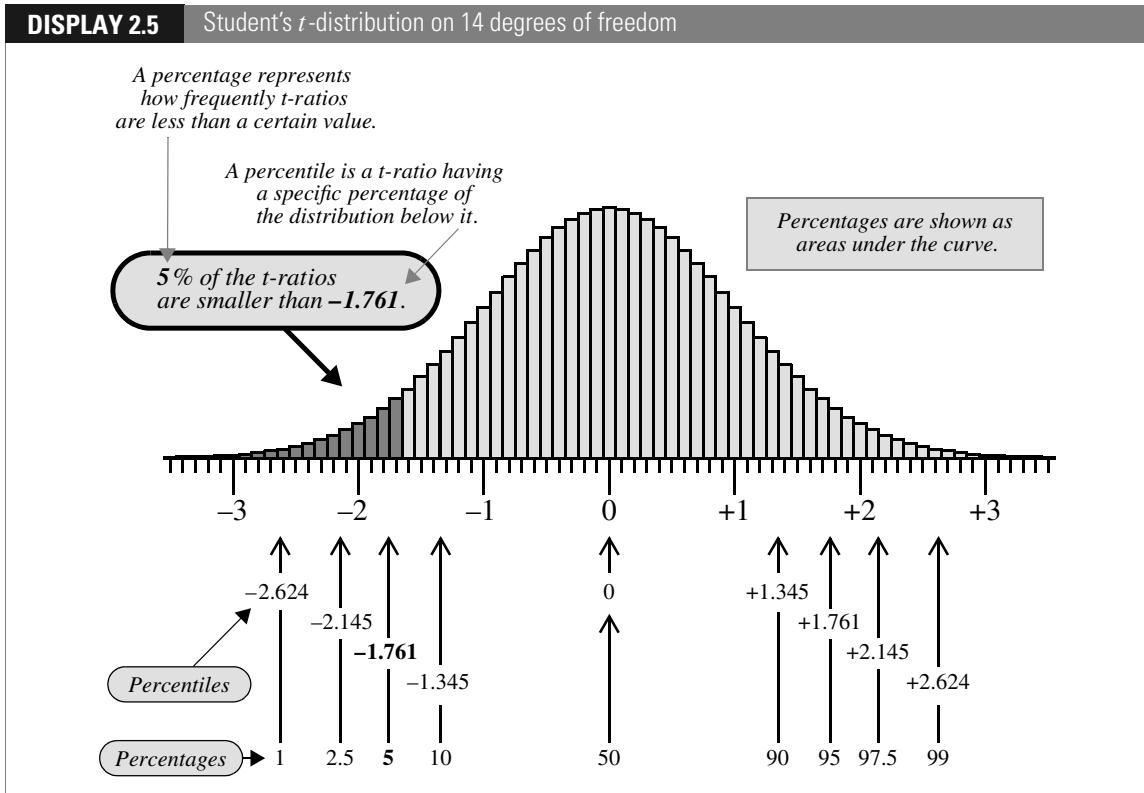
This distribution indicates the likely values for the *t*-ratio, which in turn may be used to indicate plausible values of μ .

The Paired *t*-Test for No Difference

For example, consider the question "Is it plausible, based on the data, that μ could be zero?" If μ were zero, that would imply that the *t*-ratio was

$$\text{*t*-ratio (if } \mu \text{ is zero)} = (0.199 - 0)/0.0615 = 3.236.$$

Display 2.5, however, shows that 3.236 is an unusually large value to have come from the *t*-distribution. More precisely, only 0.003 (i.e., 0.3%) of all random samples from a population in which $\mu = 0$ lead to values of the *t*-ratio as large as or larger than 3.236. (This value comes from a calculator or from a statistical computer program.) The proportion of random samples that yield *t*-ratios that are as far or farther from 0 than 3.236 in *either* direction is 0.006 (double the 0.003). So here are your choices: (a) $\mu \neq 0$; or (b) $\mu = 0$ and the random sampling resulted in a particularly nonrepresentative sample. You cannot prove that $\mu \neq 0$, but you may infer it from the rarity of the converse. This type of reasoning is the conceptual basis for testing a hypothesis about a parameter, and the measure 0.006 is the (two-sided) *p*-value based on the *t*-distribution.



A 95% Confidence Interval for the Mean

Consider also the question "What are plausible values for μ , based on the data?" This can be answered by unraveling the *t*-ratio in a slightly different way. Display 2.5 shows that the most typical *t*-ratios are near zero, with 95% of the most likely values being between -2.145 and $+2.145$. If this sample produces one of these 95% most likely *t*-ratios, then

$$-2.145 < (0.199 - \mu)/0.0615 < +2.145,$$

in which case μ is between 0.067 and 0.331 cm^3 . The interval from 0.067 to 0.331 is a 95% confidence interval for μ .

The Interpretation of a Confidence Interval

A 95% confidence interval will contain the parameter if the *t*-ratio from the observed data happens to be one of those in the middle 95% of the sampling distribution. Since 95% of all possible pairs of samples lead to such *t*-ratios, the *procedure* of constructing a 95% confidence interval is successful in capturing the parameter of interest in 95% of its applications. It is impossible to say whether it is successful or not in any particular application.

DISPLAY 2.6Calculations for the paired *t*-test and 95% confidence interval for the schizophrenia studySUMMARY STATISTICS

1

Compute differences.
Obtain their average, \bar{Y} , and
standard deviation, s .

Y_i is the difference between the left hippocampus volumes in twin i ($i=1, \dots, 15$), unaffected twin minus schizophrenic twin.
 Sample average: $\bar{Y}=0.199 \text{ cm}^3$
 Sample standard deviation: $s=0.238 \text{ cm}^3$; 14 d.f.

2

Compute the standard error of
the average: $SE(\bar{Y})=s/\sqrt{n}$
with its d.f. = $n-1$

$$SE(\bar{Y})=0.238/\sqrt{15}=0.0615 \text{ cm}^3$$

PAIRED *t*-TEST FOR THE HYPOTHESIS OF NO DIFFERENCE IN MEAN VOLUMES

3

*Compute the *t*-statistic:*
 $t=(\bar{Y}-0)/SE(\bar{Y})$

$$t\text{-statistic}=0.199/0.0615=3.236; 14 \text{ d.f.}$$

4

*Find the *p*-value (two-sided here) as the*
*proportion of *t*-ratios in a t_{n-1} distribution*
*as or more extreme than the *t*-statistic*

two-sided *p* -value = 0.006
 (from tabulated *t*-distribution with 14 d.f.)

95% CONFIDENCE INTERVAL FOR THE MEAN VOLUME DIFFERENCE

5

Find the 97.5th percentile, $t_{n-1}(0.975)$, in the
**t*-distribution with $n-1$ d.f.*

$$t_{14}(0.975)=2.145$$

(from tabulated *t*-distribution with 14 d.f.)

6

The confidence interval: $\bar{Y} \pm t_{n-1}(0.975) \times SE(\bar{Y})$

$$0.199 \pm 2.145 \times 0.0615$$

$\Rightarrow \mathbf{0.067 \text{ cm}^3 \text{ to } 0.331 \text{ cm}^3}$

A summary of the calculations for the paired *t*-test and confidence interval for the schizophrenia study are provided in Display 2.6.

2.3 A *t*-RATIO FOR TWO-SAMPLE INFERENCE

The preceding discussions provide a conceptual basis for the development of inferential tools based on the *t*-distributions. This section repeats the discussions more formally, for the structure of independent samples from two normally distributed populations.

2.3.1 Sampling Distribution of the Difference Between Two Independent Sample Averages

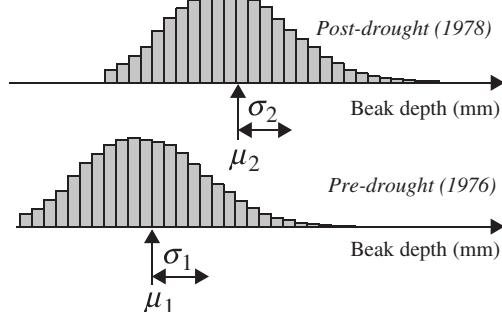
The goal is to draw conclusions about the difference in two population means from the difference in two sample averages. The latter is variable—its value depends on the particular samples that happened to have been selected—and therefore leads to an uncertain conclusion. Fortunately, mathematical theory about the sampling distribution of the difference in averages can be used to simultaneously convey the uncertain conclusion along with an indication of its uncertainty.

Display 2.7 lists the main mathematical results. The top panel shows histograms representing the two unknown population distributions and the bottom panel shows the sampling distribution for the difference in averages, meaning the probability distribution of possible values of the difference in averages that would emerge from (hypothetical) repeated sampling from the two populations. Although only a single sample is selected from each population in practice, mathematical theory based on simple random sampling reveals some useful facts about the sampling distribution, which are indicated in bubbles 1, 2, and 3.

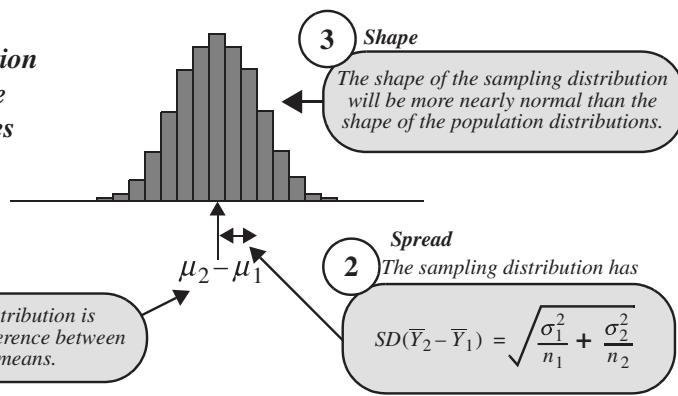
DISPLAY 2.7

Facts about the sampling distribution of the difference of averages from two independent random samples (from statistical theory)

Population Distributions



Sampling Distribution of the Difference Between Averages



The mathematical theory reveals that the shape of the sampling distribution is approximately normal and that the adequacy of the approximation improves with increasing sample sizes. It also provides a formula for the standard deviation of the sampling distribution, as shown in bubble 3 of Display 2.7. This formula isn't directly useable because it depends on the unknown population standard deviations. As in the one-sample problem, though, an estimate is obtained by replacing the unknown σ 's in the formula by estimates. As before, the resulting *estimated* standard deviation of the sampling distribution is called a *standard error*.

2.3.2 Standard Error for the Difference of Two Averages

Statisticians have devised two methods for estimating the standard deviation in the sampling distribution of the difference between two averages. Some prefer an “unequal SD” method in which the two SDs are estimated independently from the two samples. This method will be presented in Chapter 4. Others prefer an “equal SD” method in which the two SDs are assumed equal and a single estimate of the common value is made by pooling information from the two samples. This book focuses on the latter method because it is a fundamental starting point for learning about the more sophisticated tools of regression and analysis of variance, which follow in later chapters. So assume in the following that the two populations have equal standard deviations: $\sigma_1 = \sigma_2 = \sigma$.

Pooled Standard Deviation for Two Independent Samples

If the two populations have the same standard deviation, σ , then the sample standard deviations, s_1 and s_2 , are independent estimates of it. A single estimate combines the two, and such a combination is formed by averaging on the variance scale. An average is not quite right, though, since the sample variance from a larger sample should be taken more seriously than a sample variance from a smaller one. A weighted average is in order, and the best single estimate of σ^2 is the weighted average in which the individual sample variances are weighted by their degrees of freedom. The square root of this is called the *pooled estimate of standard deviation*, s_p , or, alternatively, the *pooled SD*:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}}, \quad \text{d.f.} = n_1 + n_2 - 2.$$

The number of degrees of freedom associated with this estimate is the sum of degrees of freedom from the individual estimates: $(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$.

Standard Error for the Difference

A formula for the standard deviation of the difference between averages appears in bubble 2 of Display 2.7. If the two populations have equal standard deviations, the formula simplifies. The following shows the simplified formula with the common

standard deviation replaced by the pooled estimate of it. This is the standard error for the difference in sample averages.

$$\text{SE}(\bar{Y}_2 - \bar{Y}_1) = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Display 2.8 shows the standard error calculations from the summary statistics of the two groups of finch beak depths, resulting in a value of 0.1459 mm from 176 degrees of freedom. The standard error may be interpreted as an estimate of the typical size of the deviation of $\bar{Y}_2 - \bar{Y}_1$ from the population quantity of interest, $\mu_2 - \mu_1$. This suggests that the observed difference in averages, 0.6685 mm, might depart by about 0.1459 mm from the quantity of interest. In this way, the standard error can help researchers describe the uncertainty in the uncertain conclusion from their estimate. Even better communication of the uncertainty, though, is accomplished with *p*-values and confidence intervals.

DISPLAY 2.8

Calculation of the pooled estimate of SD and the standard error for the difference between two sample averages; finch beak data

1 Summary Statistics

Group	<i>n</i>	Average (mm)	Sample SD (mm)
1: Pre-drought	89	9.4697	1.0353
2: Post-drought	89	10.1382	0.9065

2 The Pooled SD

$$\begin{aligned}
 s_p &= \sqrt{\frac{(89-1)(1.0353)^2 + (89-1)(0.9065)^2}{(89+89-2)}} \\
 &= \sqrt{\frac{166.6358}{176}} \quad \text{These are the degrees of freedom associated with the pooled SD.} \\
 &= \sqrt{0.9468} \quad \text{This is the pooled variance.} \\
 \text{Answer} \rightarrow s_p &= 0.9730 \text{ mm}
 \end{aligned}$$

3 The Standard Error

$$\begin{aligned}
 \text{SE}(\bar{Y}_2 - \bar{Y}_1) &= 0.9730 \sqrt{\frac{1}{89} + \frac{1}{89}} \\
 &= 0.1459 \text{ mm}
 \end{aligned}$$

2.3.3 Confidence Interval for the Difference Between Population Means

Inferences about the difference between population means arise from the consideration of a *t*-ratio, as in Section 2.2.3. The parameter of interest is $\mu_2 - \mu_1$. Its estimate is $\bar{Y}_2 - \bar{Y}_1$. The standard error comes from the previous section, and it has $n_1 + n_2 - 2$ degrees of freedom. In this case the *t*-ratio is $[(\bar{Y}_2 - \bar{Y}_1) - (\mu_2 - \mu_1)]/\text{SE}(\bar{Y}_2 - \bar{Y}_1)$. If the populations are normally distributed, this *t*-ratio has a *t*-distribution with $n_1 + n_2 - 2$ degrees of freedom.

For the finch beak data $\bar{Y}_2 - \bar{Y}_1 = 0.6685$ mm, $\text{SE}(\bar{Y}_2 - \bar{Y}_1) = 0.1459$ mm, and the *t*-ratio has a sampling distribution described by a *t*-distribution on 176 degrees of freedom. A statement about the likely values for the *t*-ratio from this distribution can be translated into a statement about the plausible values for $\mu_2 - \mu_1$.

A computer program that provides percentiles of *t*-distributions will reveal that 95% of values in a *t*-distribution with 176 degrees of freedom will fall between -1.9735 and $+1.9735$. A 95% confidence interval can be obtained by supposing that the actual *t*-ratio is one of these 95% in the center. The extreme *t*-ratios, -1.9735 and $+1.9735$, are now used to find corresponding endpoints for an interval on the parameter. Setting

$$-1.9735 < \frac{0.6685 - (\mu_2 - \mu_1)}{0.1459} < 1.9735$$

and solving for the parameter value yields the two interval endpoints: $0.3807 \text{ mm} < \mu_2 - \mu_1 < 0.9564 \text{ mm}$.

The preceding interval will contain $\mu_2 - \mu_1$ if the *t*-ratio for the observed samples is one of those 95% central ones, but not otherwise. The terminology “95% confidence” means that the procedure of constructing a 95% confidence interval is successful in 95% of its applications. It’s successful in the 95% of applications for which chance deals a *t*-ratio from the central part of the *t*-distribution.

The Mechanics of Confidence Interval Construction

A confidence interval with *confidence level* $100(1 - \alpha)\%$ is the following:

100(1 - α)% Confidence Limits for the Difference Between Means:

$$(\bar{Y}_2 - \bar{Y}_1) \pm t_{df}(1 - \alpha/2)\text{SE}(\bar{Y}_2 - \bar{Y}_1).$$

This formula requires that a quantity be subtracted from $\bar{Y}_2 - \bar{Y}_1$ to get the lower endpoint and that the same quantity be added to $\bar{Y}_2 - \bar{Y}_1$ to get the upper endpoint. The symbol $t_{df}(1 - \alpha/2)$ represents the $100(1 - \alpha/2)\text{th}$ percentile of the *t*-distribution on *d.f.* degrees of freedom. For example, $t_{176}(0.975)$ represents the 97.5th percentile in the *t*-distribution with 176 degrees of freedom (which is 1.9735). It may seem strange that the 97.5th percentile is desired in the calculation of a 95% confidence interval, but the 2.5th and the 97.5th percentiles are the ones that divide the middle

DISPLAY 2.9

Construction of a 95% confidence interval for the difference between the mean beak depths in 1978 and 1976

Group	<i>n</i>	Average (mm)	SD (mm)
1: Pre-drought	89	9.4697	1.0353
2: Post-drought	89	10.1382	0.9065

$$\bar{Y}_2 - \bar{Y}_1 = 10.1382 - 9.4697 = 0.6685 \text{ mm}$$

$$SE(\bar{Y}_2 - \bar{Y}_1) = 0.1459 \text{ mm}$$

$$\text{Degrees of freedom} = 89 + 89 - 2 = 176$$

$$t_{176}(0.975) = 1.9735$$

$$\text{Half-width} = (1.9735)(0.1459) = 0.2879$$

$$\text{Lower 95\% confidence limit} = 0.6685 - 0.2879 = 0.3807 \text{ mm}$$

$$\text{Upper 95\% confidence limit} = 0.6685 + 0.2879 = 0.9564 \text{ mm}$$

From Display 2.8

From tables of the
t-distribution with
176 degrees of freedom

95% of the distribution from the rest; and the 2.5th percentile is always the negative of the 97.5th. Display 2.9 summarizes the solution for the finch beak data.

Factors Affecting the Width of a Confidence Interval

There is a trade-off between the level of confidence and the width of the confidence interval. The level of confidence can be specified to be large by the user (and a high confidence level is good), but only at the expense of having a wider interval (which is bad, since the interval is less specific in answering the question of interest). If the consequences of not capturing the parameter are severe, then it might be wise to use 99% confidence intervals, even though they will be wider and, therefore, less informative. If the consequences of not capturing the parameter are minor, then a 90% interval might be a good choice. Although somewhat arbitrary, a confidence level of 95% is a conventional choice for balancing the trade-off between level of confidence and interval width.

The only way to decrease the interval width without decreasing the level of confidence (or similarly to increase the level of confidence without increasing width) is to get more data or, if possible, to reduce the size of σ . If a guess or initial estimate of σ is available, it is possible to determine the sample size needed to get a confidence interval of a certain width. This is discussed in Chapter 23.

2.3.4 Testing a Hypothesis About the Difference Between Means

To test a hypothesized value for a parameter, a *t-statistic* is formed in the same way as the *t*-ratio, but supposing the hypothesis is true. The *t*-distribution permits an evaluation of whether the *t*-statistic is a likely value for a *t*-ratio and, hence, whether the hypothesis is reasonable. The *t*-statistic for the difference in means is

$$t\text{-statistic} = \frac{(\bar{Y}_2 - \bar{Y}_1) - [\text{Hypothesized value for } (\mu_2 - \mu_1)]}{\text{SE}(\bar{Y}_2 - \bar{Y}_1)}.$$

This *t*-statistic tells how many standard errors the estimate is away from the hypothesized parameter. Its sign tells whether the estimate is above the hypothesized value (+) or below it (-).

The *p*-value, used as a measure of the credibility of the hypothesis, is the proportion of all possible *t*-ratios that are as far or farther from zero than is the *t*-statistic.

The p-value for a t-test is the probability of obtaining a t-ratio as extreme or more extreme than the t-statistic in its evidence against the null hypothesis, if the null hypothesis is correct.

The *p*-value may be based on a probability model induced by random assignment in a randomized experiment (Section 1.3.2) or on a probability model induced by random sampling from populations, as here.

If the *p*-value is small, then either the hypothesis is correct—and the sample happened to be one of those rare ones that produce such an unusual *t*-ratio—or the hypothesis is incorrect. Although it is impossible to know which of these two possibilities is true, the *p*-value indicates the probability of the first of these results and, therefore, provides a measure of credibility for that interpretation. *The smaller the p-value, the stronger is the evidence that the hypothesis is incorrect.* A large *p*-value implies that the study is not capable of excluding the null hypothesis as a possible explanation for how the data turned out. A possible wording in this case is “the data are consistent with the hypothesis being true.” It is wrong to conclude that the null hypothesis *is true*.

One-Sided and Two-Sided *p*-Values

In the finch beak example the *t*-statistic for testing the hypothesis of “no difference” in population means is 4.583. The proportion of *t*-ratios that are as far or farther from zero than 4.583 is found from the percentiles of the t_{176} distribution. The proportion of *t*-ratios *greater than or equal* to 4.583 is minuscule, < 0.00001 . The *t*-ratios *as far or farther from zero* than 4.583 are those less than or equal to -4.583 and those greater than or equal to 4.583, and this proportion is twice the proportion greater than 4.583 (because *t*-distributions are symmetric), but still minuscule.

The proportion of *t*-ratios farther from zero in one specified direction is referred to as a *one-sided p-value*. The proportion of *t*-ratios farther from zero than the *t*-statistic, either positively or negatively, is a *two-sided p-value*.

The choice of one-sided or two-sided depends on how specifically the researcher wishes to declare the alternatives to the hypothesis of equal means. In the finch beak study, the Grants surmised that the response of the species to the sole availability

of large, tough seeds would not merely be a change in mean beak sizes but more specifically a tendency toward larger beak sizes. If their intention is to report a conclusion about the evidence that the mean beak size *increased*, regardless of how the data actually turn out, then a one-sided p -value is in order.

Much has been made of whether to report one-sided or two-sided p -values. There are some situations where a one-sided p -value is appropriate, some where a two-sided p -value is appropriate, and many where it is not at all clear which is appropriate. Since the two provide equivalent measures of evidence against the hypothesis of equal means (that is, the two-sided p -value is simply twice the one-sided p -value), the distinction is not terribly important; a reader may convert one to other. There is only one absolute when it comes to reporting: *always report whether the p-value is one- or two-sided*.

2.3.5 The Mechanics of p -Value Computation

The steps required to compute a p -value for the test of the hypothesis that $\mu_2 - \mu_1 = D$ (a specified value, like 0) are as follows.

1. Compute the estimate, $\bar{Y}_2 - \bar{Y}_1$, its standard error, and the degrees of freedom.
2. Compute the t -statistic: $t = [(\bar{Y}_2 - \bar{Y}_1) - D]/SE(\bar{Y}_2 - \bar{Y}_1)$.
3. Determine the proportion, P , of t -ratios that are less than the t -statistic, using a statistical computer program with t -distribution percentiles with the appropriate degrees of freedom.
4. Determine the p -value based on the proportion, P , and the alternatives of interest. (a) For one-sided alternatives of the form $\mu_2 - \mu_1 > D$, t -ratios larger than t are more extreme, so the one-sided p -value is $1 - P$. (b) For the one-sided alternatives of the form $\mu_2 - \mu_1 < D$, t -ratios smaller than t are more extreme, so the one-sided p -value is P . (c) For two-sided alternatives of the form $\mu_2 - \mu_1 \neq D$, t -ratios that are larger in magnitude than t are more extreme, so the two-sided p -value is $2P$ if $P < 0.5$ or $2(1 - P)$ if $P > 0.5$.
5. Report the hypothesis, the p -value, and whether it is one- or two-sided.

An illustration of this procedure for the finch beak data is shown in Display 2.10.

2.4 INFERENCES IN A TWO-TREATMENT RANDOMIZED EXPERIMENT

Probability models for randomized experiments (Section 1.3.1) are spawned by the chance mechanisms used to assign subjects to treatment groups. Probability models for random sampling (Section 1.4.1) are spawned by the chance mechanisms used to select units from real, finite populations. Chapter 2 has thus far discussed inference procedures whose motivation stems from considerations of random sampling from populations that are conceptual, infinite, and normally distributed. While there seems to be a considerable difference between the situations, it turns out that the t -distribution uncertainty measures discussed in this chapter are useful approximations to both the randomization and the random sampling uncertainty measures for a wide range of problems. The practical consequence is that t -tools

DISPLAY 2.10

The *t*-test for the hypothesis that the post-drought and pre-drought beak depth means are equal

Group	<i>n</i>	Average (mm)	SD (mm)
1: Pre-drought	89	9.4697	1.0353
2: Post-drought	89	10.1382	0.9065

$\bar{Y}_2 - \bar{Y}_1 = 10.1382 - 9.4697 = 0.6685 \text{ mm}$

$SE(\bar{Y}_2 - \bar{Y}_1) = 0.1459 \text{ mm}$

Degrees of freedom = $89 + 89 - 2 = 176$

$t\text{-statistic} = \frac{0.6685 - 0.0}{0.1459} = 4.583$

$P = 0.999996$

One-sided *p*-value = 0.000004

From Display 2.8

Hypothesized difference

*From tables of the *t*-distribution with 176 degrees of freedom:*

are used for many situations that do not conform to the strict model upon which the *t*-tools are based, including data from randomized experiments. Conclusions from randomized experiments, however, are phrased in the language of treatment effects and causation, rather than differences in population means and association.

2.4.1 Approximate Uncertainty Measures for Randomized Experiments

Although theoretically motivated by random samples from normal populations, the two-sample *t*-test can also be applied to data from a two-group randomized experiment. Reconsideration of the creativity study (Section 1.1.1) illustrates the procedure. In the intrinsic group, the average of 24 scores is 19.88, and the SD is 4.44. In the extrinsic group, the average of 23 scores is 15.74, and the SD is 5.25. The pooled estimate of the standard deviation is 4.85, and the standard error for the difference between averages is 1.42 with 45 degrees of freedom. The difference between average scores is 4.14 points.

Hypothesis Test of No Treatment Effect

The *t*-statistic, $t = (4.14 - 0)/1.42 = 2.92$, can be used to test the hypothesis of no treatment effect in exactly the same way it would be used to test equal population means from two random samples. The probability in a *t*-distribution on 45 degrees of freedom to the right of 2.92 is 0.0027, so the one-sided *p*-value for the alternative of a positive difference is 0.0027, and the two-sided *p*-value is 0.0054. The conclusion is phrased in terms of the additive treatment effect model: “This experiment provides strong evidence that receiving the intrinsic questionnaire caused a higher creativity score (one-sided *p*-value = 0.0027).”

Confidence Interval for a Treatment Effect

Construction of a confidence interval for an additive treatment effect δ is precisely the same as for the difference between population means, $\mu_2 - \mu_1$. The 97.5th percentile in the t -distribution with 45 degrees of freedom is 2.014, so the interval half-width is $(2.014)(1.42) = 2.85$. The interval runs from $4.14 - 2.85 = 1.29$ to $4.14 + 2.85 = 6.99$ points.

Compare this construction with one based on the randomization procedure itself (Section 1.3). The randomization-based procedure relies on a relationship between testing and confidence intervals: *Any hypothesized parameter value should be included or excluded from a $100(1-\alpha)\%$ confidence interval according to whether its test yields a two-sided p -value that is greater than or less than α .* Accordingly, to construct a 95% confidence interval for the treatment effect δ , include only those values of δ which, when tested, have two-sided p -values greater than 0.05.

To determine whether $\delta = 5$ should be included in a 95% confidence interval, one must consider the randomization model for $\delta = 5$. If this is the correct value, subtracting 5 from the scores of all persons in the intrinsic group reconstructs the scores they would have had if placed in the extrinsic group. Now all 47 scores are homogeneous, so a randomization test should conclude there is no difference. Perform the randomization test. If the two-sided p -value exceeds (or equals) 0.05, include 5 in the interval. Otherwise, leave it out. That settles the issue for $\delta = 5$, but the limits of the interval must be found by repeating this process to find the smallest and largest δ for which the two-sided p -value is greater than or equal to 0.05.

Approximation of the Randomization Distribution of the t -Statistic

The t -based p -values and confidence intervals are only approximations to the correct p -values and confidence intervals from randomization distributions. To see how good the approximation is in the creativity study, the analysis of Chapter 1 was modified by considering the randomization distribution of the t -statistic rather than the difference in sample averages. A histogram of the t -statistics from 500,000 random regroupings appears in Display 2.11, along with the approximating t -distribution. The observed t -statistic (2.92) was exceeded by only 1,298 of the 500,000 random regroupings, giving an estimated one-sided p -value of $1,298/500,000 = 0.0026$. The approximation based on the t -distribution (0.0027) is quite good.

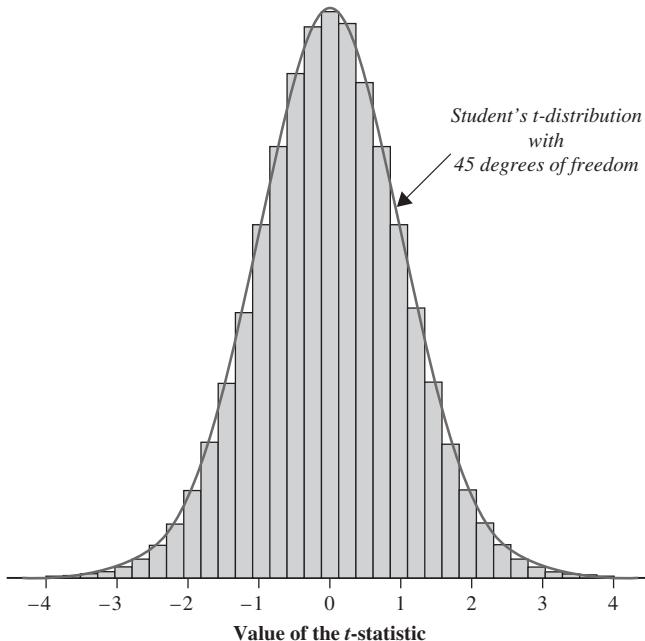
2.5 RELATED ISSUES

2.5.1 Interpretation of p -Values

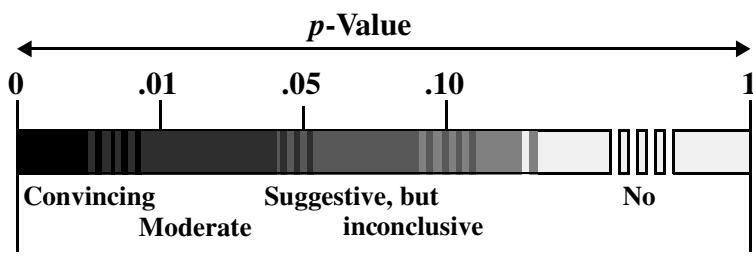
A small p -value like 0.005 means either that the null hypothesis is correct—and the randomization or random sampling led by chance to a rare outcome where the t -ratio is quite far from zero—or that the hypothesis is incorrect. The smaller the p -value, the greater is the evidence that the second explanation is the correct one.

DISPLAY 2.11

A histogram of *t*-statistics from 500,000 random regroupings of the creativity study data with an approximating *t*-distribution curve, demonstrating that the easier-to-use *t*-tools can give very good approximations to randomization test inferences, even though they are derived for a quite different model—random samples from two normal populations

**DISPLAY 2.12**

Interpreting the size of a *p*-value



Is there evidence of a difference?

How small is small? It is difficult and unwise to decide on absolute cutoff points for believability to be applied in all situations. Display 2.12 represents a starting point for interpreting *p*-values.

P-values can be comprehended by comparing them to events whose probabilities are more familiar. For example, at what point does a person flipping a series

of heads begin to doubt that the coin is fair? It is not terribly unlikely to get four heads in a row. The probability of doing so with a fair coin is 0.0625. At five heads in a row one might start to get a bit curious. The chance of this, if the coin is fair, is 0.03125. When the sixth toss is also heads, one may start to question the integrity of the coin, even though it is possible that a fair coin could turn up heads six times in a row (with probability 0.015625). Ten heads in a row is convincing evidence that the coin is not of standard issue. The probability of this event, if the chance of heads were in fact one-half, is 0.0009766.

It is tempting to think of a *p*-value as the probability of the null hypothesis being correct, but this interpretation is technically incorrect and potentially misleading. The hypothesis is or is not correct, and there is no probability associated with that. The probability arises from the uncertainty in the data. So the best technical description is the precise definition (see Section 2.3.4), clumsy as it may sound.

2.5.2 An Example of Confidence Intervals

In 1915 Albert Einstein published four papers in the proceedings of the Prussian Academy of Sciences laying the foundations of general relativity and describing some of its consequences. A paper establishing the field equation for gravity predicted that the arc of deflection of light passing the sun would be twice the angle predicted by Newton's gravity theory. Half the predicted deflection comes directly from Newton's calculations, and the other half comes from the curvature of space near the sun relative to space far away. This is represented by the equation

$$\Delta = (1/2)(1 + \gamma) \frac{1.75}{d},$$

where Δ is the deflection of light, d is the distance of the closest approach of the ray to the sun (in solar radii), and γ is the parameter describing space curvature.

The parameter γ , which is predicted by Einstein's general relativity theory to be 1 and by Newtonian physics to be 0, was estimated in 1919 by British astronomers during a total solar eclipse. Since then, measurements have been repeated many times, under various measurement conditions. The efforts are summarized in Display 2.13. (Data from C. M. Will, "General Relativity at 75: How Right Was Einstein?" *Science* 250 (November 9, 1990): 770–75.)

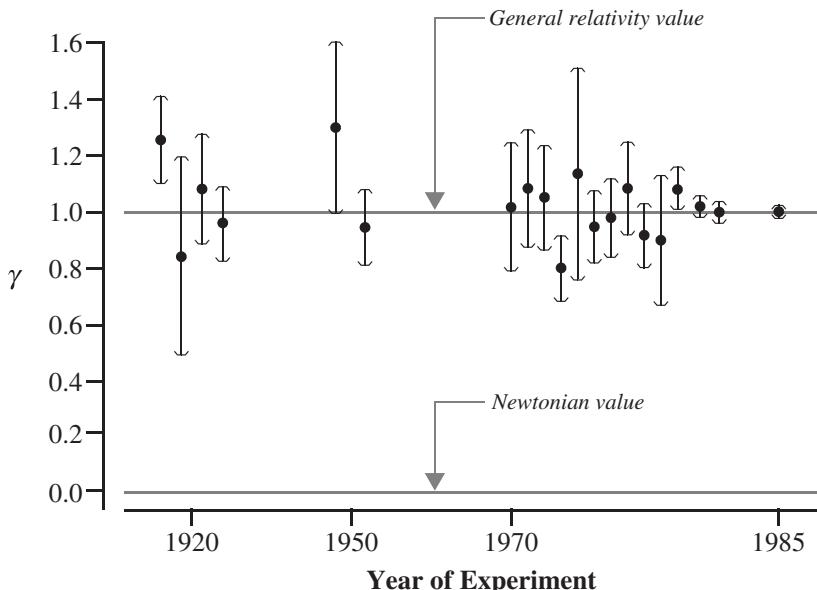
The confidence intervals around the estimates in Display 2.13 reflect uncertainty due to measurement errors. (The actual confidence levels are not given and are not important for this illustration.) After the first relatively crude attempts to measure γ , little improvement was made until the late 1960s and the discovery of quasars. Measurements of light from quasar groups passing near the sun led to dramatic improvement in accuracy, as evident in the narrower intervals with later years.

A Note About the Accumulation of Evidence

This example shows that theories must withstand continual challenges from skeptical scientists. The essence of scientific theory is the ability to predict future

DISPLAY 2.13

Estimates and confidence intervals for γ , the deflection of light around the sun, from 20 experiments



outcomes. Experimental results are typically uncertain. So the fact that some intervals fail to include the value $\gamma = 1$ is not taken to disprove general relativity, but neither would it prove general relativity right if all the intervals did include $\gamma = 1$. When a theory's predictions are consistently denied by a series of experiments—such as the Newtonian prediction of $\gamma = 0$ in this example—scientists agree that the theory is not adequate.

2.5.3 The Rejection Region Approach to Hypothesis Testing

Not long ago, statisticians took a *rejection region* approach to testing hypotheses. A *significance level* of 0.05, say, was selected in advance, and a *p*-value less than 0.05 called for rejecting the hypothesis at the significance level 0.05; otherwise, the hypothesis was accepted, or more correctly, not rejected. Thus *p*-values of 0.048 and 0.0001 both lead to rejection at the 0.05 level, even though they supply vastly different degrees of evidence. On the other hand, *p*-values of 0.049 and 0.051 lead to different conclusions even though they provide virtually identical evidence against the hypothesis. Although important for leading to advances in the theory of statistics, the rejection region approach has largely been discarded for practical applications and *p*-values are reported instead. *P*-values give the reader more information for judging the significance of findings.

2.6 SUMMARY

Many research questions can be formulated as comparisons of two population distributions. Comparison of the distributions' centers effectively summarizes the difference between the parameters of interest when the populations have the same variation and general shape. This chapter concentrated on the difference in means, $\mu_2 - \mu_1$, which is estimated by the difference in sample averages.

The statistical problem is to assess the uncertainty associated with the difference between the estimate (the difference in sample averages) and the parameter (the difference in population means). The sampling distribution of an estimate is the key to understanding the uncertainty. It is represented as a histogram of values of the estimate for every possible sample that could have been selected.

Often with fairly large samples, a sampling distribution has a normal shape. A normal sampling distribution is specified by its mean and its standard deviation. When the populations have common standard deviation σ the difference in sample averages has a sampling distribution with mean $\mu_2 - \mu_1$ and standard deviation

$$\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

This could be used to describe uncertainty except that it involves the unknown σ —the common standard deviation in the two populations. In practice, σ is replaced by its best estimate from the data—the pooled standard deviation, s_p , having $n_1 + n_2 - 2$ degrees of freedom. The estimated standard deviation of the sampling distribution is called the standard error.

The standard error alone, however, does not entirely describe the uncertainty in an estimate. More precise statements can be made by using the t -ratio, which has a Student's t -distribution as its sampling distribution (if the ideal normal model applies). This leads directly to confidence intervals and p -values as statistical tools for answering the questions of interest. The confidence interval provides a range of likely values for the parameter, and the confidence level is interpreted as the frequency with which the interval construction procedure gives the right answer. For testing whether a particular hypothesized number could be the unknown parameter, the t -statistic is formed by substituting the hypothesized value for the true value in the t -ratio. The p -value is the chance of getting as extreme or more extreme t -ratios than the t -statistic, and it is interpreted as a measure of the credibility of the hypothesized value.

2.7 EXERCISES

Conceptual Exercises

1. **Finch Beak Data.** Explain why the finch beak study is considered an observational study.

2. For comparing two population means when the population distributions have the same standard deviation, the standard deviation is sometimes referred to as a nuisance parameter. Explain why it might be considered a nuisance.
3. True or false? If a sample size is large, then the shape of a histogram of the sample will be approximately normal, even if the population distribution is not normal.
4. True or false? If a sample size is large, then the shape of the sampling distribution of the average will be approximately normal, even if the population distribution is not normal.
5. Explain the relative merits of 90% and 99% levels of confidence.
6. What is wrong with the hypothesis that $\bar{Y}_2 - \bar{Y}_1$ is 0?
7. In a study of the effects of marijuana use during pregnancy, measurements on babies of mothers who used marijuana during pregnancy were compared to measurements on babies of mothers who did not. (Data from B. Zuckerman et al., "Effects of Maternal Marijuana and Cocaine Use on Fetal Growth," *New England Journal of Medicine* 320(12) (March 1989): 762–68.) A 95% confidence interval for the difference in mean head circumference (nonuse minus use) was 0.61 to 1.19 cm. What can be said from this statement about a *p*-value for the hypothesis that the mean difference is zero?
8. Suppose the following statement is made in a statistical summary: "A comparison of breathing capacities of individuals in households with low nitrogen dioxide levels and individuals in households with high nitrogen dioxide levels indicated that there is no difference in the means (two-sided *p*-value = 0.24)." What is wrong with this statement?
9. What is the difference between (a) the mean of Y and the mean of \bar{Y} ? (b) the standard deviation of Y and the standard deviation of \bar{Y} ? (c) the standard deviation of \bar{Y} and the standard error of \bar{Y} ? (d) a *t*-ratio and a *t*-statistic?
10. Consider blood pressure levels for populations of young women using birth control pills and young women not using birth control pills. A comparison of these two populations through an observational study might be consistent with the theory that the pill elevates blood pressure levels. What tool is appropriate for addressing whether there is a difference between these two populations? What tool is appropriate for addressing the likely size of the difference?
11. The data in Display 2.14 are survival times (in days) of guinea pigs that were randomly assigned either to a control group or to a treatment group that received a dose of tubercle bacilli. (Data from K. Doksum, "Empirical Probability Plots and Statistical Inference for Nonlinear Models in the Two-Sample Case," *Annals of Statistics* 2(1974): 267–77.) (a) Why might the additive treatment effect model (introduced in Section 1.3.1) be inappropriate for these data? (b) Why might the ideal normal model with equal spread be an inadequate approximation?

Computational Exercises

12. **Marijuana Use During Pregnancy.** For the birth weights of babies in two groups, one born of mothers who used marijuana during pregnancy and the other born of mothers who did not (see Exercise 7), the difference in sample averages (nonuser mothers minus user mothers) was 280 grams, and the standard error of the difference was 46.66 grams with 1,095 degrees of freedom. From this information, provide the following: (a) a 95% confidence interval for $\mu_2 - \mu_1$, (b) a 90% confidence interval for $\mu_2 - \mu_1$, and (c) the two-sided *p*-value for a test of the hypothesis that $\mu_2 - \mu_1 = 0$.
13. **Fish Oil and Blood Pressure.** Reconsider the changes in blood pressures for men placed on a fish oil diet and for men placed on a regular oil diet, from Chapter 1, Exercise 12. Do the following steps to compare the treatments.

DISPLAY 2.14

Lifetimes of guinea pigs in two treatment groups

	36,18	0	76,93,97	
	91,89,87,86,52,50			
49,20,19,18,15,14,14,08,02		1	07,08,13,14,19,36,38,39	
89,78,73,67,67,66,65,60			52,54,54,60,64,64,66,68,78,79,81,81,83,85,94,98	
16,12,09		2	12,13,16,20,25,25,44	
92,79,78,73			53,56,59,65,68,70,83,89,91	
41		3	11,15,26,26	
<i>Control</i>	82,80,67,55		61,73,73,76,97,98	
(n=64)	46,32,21,21	4	06	
	74,63,55		59,66	<i>Received bacilli</i>
	46,45,05	5	92,98	(n=58)
	90,76,69			
41,38,37,34,21,08,07,03		6		
	88,85,63,50			
	35,25	7		

Legend: | 5 | 98 represents 598 days

- (a) Compute the averages and the sample standard deviations for each group separately.
- (b) Compute the pooled estimate of standard deviation using the formula in Section 2.3.2.
- (c) Compute $SE(\bar{Y}_2 - \bar{Y}_1)$ using the formula in Section 2.3.2.
- (d) What are the degrees of freedom associated with the pooled estimate of standard deviation? What is the 97.5th percentile of the t -distribution with this many degrees of freedom?
- (e) Construct a 95% confidence interval for $\mu_2 - \mu_1$ using the formula in Section 2.3.3.
- (f) Compute the t -statistic for testing equality as shown in Section 2.3.5.
- (g) Find the one-sided p -value (as evidence that the fish oil diet resulted in greater reduction of blood pressure) by comparing the t -statistic in (f) to the percentiles of the appropriate t -distribution (by reading the appropriate percentile from a computer program or calculator).

14. Fish Oil and Blood Pressure. Find the 95% confidence interval and one-sided p -value asked for in Exercise 13(e) and (g) but use a statistical computer package to do so.

15. Auto Exhaust and Lead Concentration in Blood. Researchers took independent random samples from two populations of police officers and measured the level of lead concentration in their blood. The sample of 126 police officers subjected to constant inhalation of automobile exhaust fumes in downtown Cairo had an average blood level concentration of $29.2 \text{ } \mu\text{g/dl}$ and an SD of $7.5 \text{ } \mu\text{g/dl}$; a control sample of 50 police officers from the Cairo suburb of Abbassia, with no history of exposure, had an average blood level concentration of $18.2 \text{ } \mu\text{g/dl}$ and an SD of $5.8 \text{ } \mu\text{g/dl}$. (Data from A.-A. M. Kamal, S. E. Eldamaty, and R. Faris, "Blood Lead Level of Cairo Traffic Policemen," *Science of the Total Environment* 105(1991): 165–70.) Is there convincing evidence of a difference in the population averages?

16. Motivation and Creativity. Verify the statements made in the summary of statistical findings for the Motivation and Creativity Data (Section 1.1.1) by analyzing the data on the computer.

17. Sex Discrimination. Verify the statements made in the summary of statistical findings for the Sex Discrimination Data (Section 1.1.2) by analyzing the data on the computer.

18. The Grants' Finch Complete Beak Data. The data file ex0218 contains the beak depths (in mm) of all 751 finches captured by Peter and Rosemary Grant in 1976 and all 89 finches captured in 1978 (as described in Section 2.1.1). Use a statistical computer program for parts a–d: (a) Draw side-by-side box plots of the two groups of beak depths. (b) Use the two-sample *t*-test on these data to find the one-sided *p*-value for a test of the hypothesis of no difference in means against the alternative that the mean in 1978 is larger. (c) What is the two-sided *p*-value from the *t*-test? (d) Provide an estimate and a 95% confidence interval for the amount by which the 1978 mean exceeds the 1976 mean. (e) What is it about the finches in the two populations that might make you question the validity of the independence assumption upon which the two-sample *t*-test is derived?

19. Fish Oil and Blood Pressure. Reconsider the fish oil and blood pressure data of Chapter 1, Exercise 12. Since the measurements are the reductions in blood pressure for each man, it is of interest to know whether the mean reduction is zero for each group. For the regular oil diet group do the following:

- (a) Compute the average and the sample standard deviation. What are the degrees of freedom associated with the sample standard deviation, s_2 ?
- (b) Compute the standard error for the average from this group: $SE(\bar{Y}_2) = s_2 / \sqrt{n_2}$.
- (c) Construct a 95% confidence interval for μ_2 as $\bar{Y}_2 + t_d(.975)SE(\bar{Y}_2)$, where d is the degrees of freedom associated with s_2 .
- (d) For the hypothesis that μ_2 is zero, construct the *t*-statistic $\bar{Y}_2/SE(\bar{Y}_2)$. Find the two-sided *p*-value as the proportion of values from a t_d -distribution farther from 0 than this value.

20. Fish Oil and Blood Pressure (One-Sample Analysis). Repeat Exercise 19 for the group of men who were given the fish oil diet and then answer these questions: Is there any evidence that the mean reduction for this group is different from zero? What is the typical reduction in blood pressure expected from this type of diet (for individuals like these men)? Provide a 95% confidence interval.

Data Problems

21. Bumpus Natural Selection Data. In 1899, biologist Hermon Bumpus presented as evidence of natural selection a comparison of numerical characteristics of moribund house sparrows that were collected after an uncommonly severe winter storm and which had either perished or survived as a result of their injuries. Display 2.15 shows the length of the humerus (arm bone) in inches for 59 of these sparrows, grouped according to whether they survived or perished. Analyze these data to summarize the evidence that the distribution of humerus lengths differs in the two populations. Write a brief paragraph of statistical conclusion, using the ones in Section 2.1 as a guide, including a

DISPLAY 2.15

Humerus lengths of moribund male house sparrows measured by Hermon Bumpus, grouped according to survival status

Humerus Lengths (inches) of 35 Males That Survived

0.687, 0.703, 0.709, 0.715, 0.721, 0.723, 0.723, 0.726, 0.728, 0.728, 0.728, 0.729, 0.730, 0.730, 0.733, 0.733, 0.735, 0.736, 0.739, 0.741, 0.741, 0.741, 0.741, 0.743, 0.749, 0.751, 0.752, 0.752, 0.755, 0.756, 0.766, 0.767, 0.769, 0.770, 0.780

Humerus Lengths (inches) of 24 Males That Perished

0.659, 0.689, 0.702, 0.703, 0.709, 0.713, 0.720, 0.720, 0.726, 0.726, 0.729, 0.731, 0.736, 0.737, 0.738, 0.738, 0.739, 0.743, 0.744, 0.745, 0.752, 0.752, 0.754, 0.765

graphical display, a conclusion about the degree of evidence of a difference, and a conclusion about the size of the difference in distributions.

22. Male and Female Intelligence. Males and females tend to exhibit different types of intelligence. Although there is substantial variability between individuals of the same gender, males on average tend to perform better at navigational and spatial tasks, and females tend to perform better at verbal fluency and memory tasks. This is not a controversial conclusion. Some researchers, however, ask whether males and females differ, on average, in their overall intelligence, and that *is* controversial because any single intelligence measure must rely on premises about the types of intelligence that are important. Even if researchers don't make a subjective judgment about a type of intelligence being tested, they are constrained by the available tools for measuring intelligence. Mathematical knowledge is easy to test, for example, but wisdom, creativity, practical knowledge, and social skill are not.

Display 2.16 shows the first five rows of a data set with several intelligence test scores for random samples of 1,306 American men and 1,278 American women between the ages of 16 and 24 in 1981. The column labeled AFQT shows the percentile scores on the Armed Forces Qualifying Test, which is designed for evaluating the suitability of military recruits but which is also used by researchers as a general intelligence test. The AFQT score is a combination of scores from four component tests: word knowledge, paragraph comprehension, arithmetic reasoning, and mathematical knowledge. The data set represented in Display 2.16 includes each individual's score on these components. (The overall AFQT score reported here, officially called AFQT89, is based on a nontrivial combination of the component scores)

DISPLAY 2.16

Armed Forces Qualifying Test (AFQT) score percentile and component test scores in arithmetic reasoning, word knowledge, paragraph comprehension, and mathematical knowledge, for 1,278 women and 1,306 men in 1981; first 5 of 2,584 rows

Gender	Arith	Word	Parag	Math	AFQT
male	19	27	14	14	70.3
female	23	34	11	20	60.4
male	30	35	14	25	98.3
female	30	35	13	21	84.7
female	13	30	11	12	44.5

Analyze the data to summarize the evidence of differences in male and female distributions of AFQT scores. Do they differ? By how much do they differ? Also answer these two questions for each of the four component test scores. Write a statistical report that includes graphical displays and statistical conclusions (like those in the case studies of Section 2.1), and a section of details upon which the conclusions were based (such as a listing of the computer output showing the results of two-sample t -tests and confidence intervals).

Notes about the data: Although these are random samples of American men and women between the ages of 16 and 24 in 1981, they are not simple random samples. The data come from the National Longitudinal Study of Youth (NLSY), which used variable probability sampling (see Section 1.5.4). To estimate the means of the larger populations, more advanced techniques are appropriate. For comparing male and female distributions, the naive approach based on random sampling is not likely to be misleading. These data come from the National Longitudinal Survey of Youth, U.S. Bureau of Labor Statistics, <http://www.bls.gov/nls/home.htm> (May 8, 2008). Rows with missing values of variables, including variables used in related problems in other chapters, have been omitted.

23. Speed Limits and Traffic Fatalities. The National Highway System Designation Act was signed into law in the United States on November 28, 1995. Among other things, the act abolished the federal mandate of 55-mile-per-hour maximum speed limits on roads in the United States and permitted states to establish their own limits. Of the 50 states (plus the District of Columbia), 32 increased their speed limits either at the beginning of 1996 or sometime during 1996. Shown in Display 2.17 are the percentage changes in interstate highway traffic fatalities from 1995 to 1996. What evidence is there that the percentage change was greater in states that increased their speed limits? How much of a difference is there? Write a brief statistical report detailing the answers to these questions. (Data from “Report to Congress: The Effect of Increased Speed Limits in the Post-NMSL Era,” National Highway Traffic Safety Administration, February, 1998; available in the reports library at <http://www-fars.nhtsa.dot.gov/>.)

DISPLAY 2.17				
Number of traffic fatalities in 50 U.S. states and the District of Columbia, and status of speed limit change in the state (retained 55 mph limit or increased speed limit); first 5 of 51 rows				
State	Fatalities1995	Fatalities1996	PctChange	SpeedLimit
Alabama	1114	1146	2.87	Inc
Alaska	87	81	-6.9	Ret
Arizona	1035	994	-3.96	Inc
Arkansas	631	615	-2.54	Inc
California	4192	3989	-4.84	Inc

Answers to Conceptual Exercises

1. The birds were *observed* to be in the 1976 or 1978 groups, not *assigned* by the researchers.
2. There is rarely any direct interest in the standard deviation, but it must be estimated in order to clear up the picture regarding means.
3. False.
4. True.
5. There is more confidence that a 99% interval contains the parameter of interest, but the extra confidence comes at the price of the interval being larger and therefore less informative about specific likely values.
6. The hypothesis must be about the population means. A hypothesis must be about the value of an *unknown* parameter. The value of a statistic will be known when the data are collected and analyzed.
7. It is less than 0.05.
8. The statement implies that the null hypothesis is accepted as true. It should be worded as, for example, the data are consistent with the hypothesis that there is no difference. (This issue is partly one of semantics, but it is still important to understand the distinction being made.)
9. (a) The mean of Y is the mean in the population of all individual measurements, and the mean of \bar{Y} is the mean of the sampling distribution of the sample mean. With random sampling, the two have the same value μ .
- (b) The standard deviation of Y is the standard deviation among all observations in the population, and the standard deviation of \bar{Y} is the standard deviation in the sampling distribution

of the sample average. The two are related, but not the same: if the standard deviation of Y is denoted by σ , then the standard deviation of \bar{Y} is σ/\sqrt{n} .

- (c) The standard error is an estimate of the standard deviation in the sampling distribution, obtained by replacing the unknown population standard deviation in the formula by the known sample standard deviation.
- (d) The t -ratio is the ratio of the difference between the estimate and the parameter to the standard error of the estimate. It involves the parameter, so you do not generally know what it is. The t -statistic is a trial value of the t -ratio, obtained when a hypothesized value of the parameter is used in place of the actual value.

10. A p -value. A confidence interval.

11. (a) Because the spread of the stem-and-leaf plot is larger for the control group than for the treatment group, it does not appear that the effect of treatment was simply to add a certain number of days onto the lives of every guinea pig. It may have added days for those that would not have lived long anyway, and subtracted days from those that would have lived a long time. (b) The equal variation assumption does not appear to be appropriate.

A Closer Look at Assumptions

Although statistical computer programs faithfully supply confidence intervals and p -values whenever asked, the human data analyst must consider whether the assumptions on which use of the tools is based are met, at least approximately. In this regard, an important distinction exists between the mathematical assumptions on which use of t -tools is exactly justified and the broader conditions under which such tools work quite well.

The mathematical assumptions—such as those of the model for two independent samples from normal populations with the same standard deviation—are never strictly met in practice, nor do they have to be. The two-sample t -tools are often valid even if the population distributions are nonnormal or the standard deviations are unequal. An understanding of the broader conditions, provided by advanced statistical theory and computer simulation, is needed to evaluate the appropriateness of the tools for a particular problem. After checking the actual conditions with graphical displays of the data, the data analyst may decide to use the standard tools, use them but apply the label “approximate” to the inferences, or choose an alternative approach.

An effective alternative is to apply the standard tools after transforming the data. A transformation is useful if the tools are appropriate for the conditions of the transformed data and if the questions of interest are answerable on the new scale. A particularly important transformation is the logarithm, which permits a convenient description of a multiplicative effect.

3.1 CASE STUDIES

3.1.1 Cloud Seeding to Increase Rainfall—A Randomized Experiment

The data in Display 3.1 were collected in southern Florida between 1968 and 1972 to test a hypothesis that massive injection of silver iodide into cumulus clouds can lead to increased rainfall. (Data from J. Simpson, A. Olsen, and J. Eden, “A Bayesian Analysis of a Multiplicative Treatment Effect in Weather Modification,” *Technometrics* 17 (1975): 161–66.)

DISPLAY 3.1 Rainfall (acre-feet) for days with and without cloud seeding									
Rainfall from Unseeded Days ($n=26$)									
1,202.6	830.1	372.4	345.5	321.2	244.3	163.0	147.8	95.0	
87.0	81.2	68.5	47.3	41.1	36.6	29.0	28.6	26.3	
26.0	24.4	21.4	17.3	11.5	4.9	4.9	1.0		
Rainfall from Seeded Days ($n=26$)									
2,745.6	1,697.1	1,656.4	978.0	703.4	489.1	430.0	334.1	302.8	
274.7	274.7	255.0	242.5	200.7	198.6	129.6	119.0	118.3	
115.3	92.4	40.6	32.7	31.4	17.5	7.7	4.1		

On each of 52 days that were deemed suitable for cloud seeding, a random mechanism was used to decide whether to seed the target cloud on that day or to leave it unseeded as a control. An airplane flew through the cloud in both cases, since the experimenters and the pilot were themselves unaware of whether on any particular day the seeding mechanism in the plane was loaded or not (that is, they were *blind* to the treatment). Precipitation was measured as the total rain volume falling from the cloud base following the airplane seeding run, as measured by radar. Did cloud seeding have an effect on rainfall in this experiment? If so, how much?

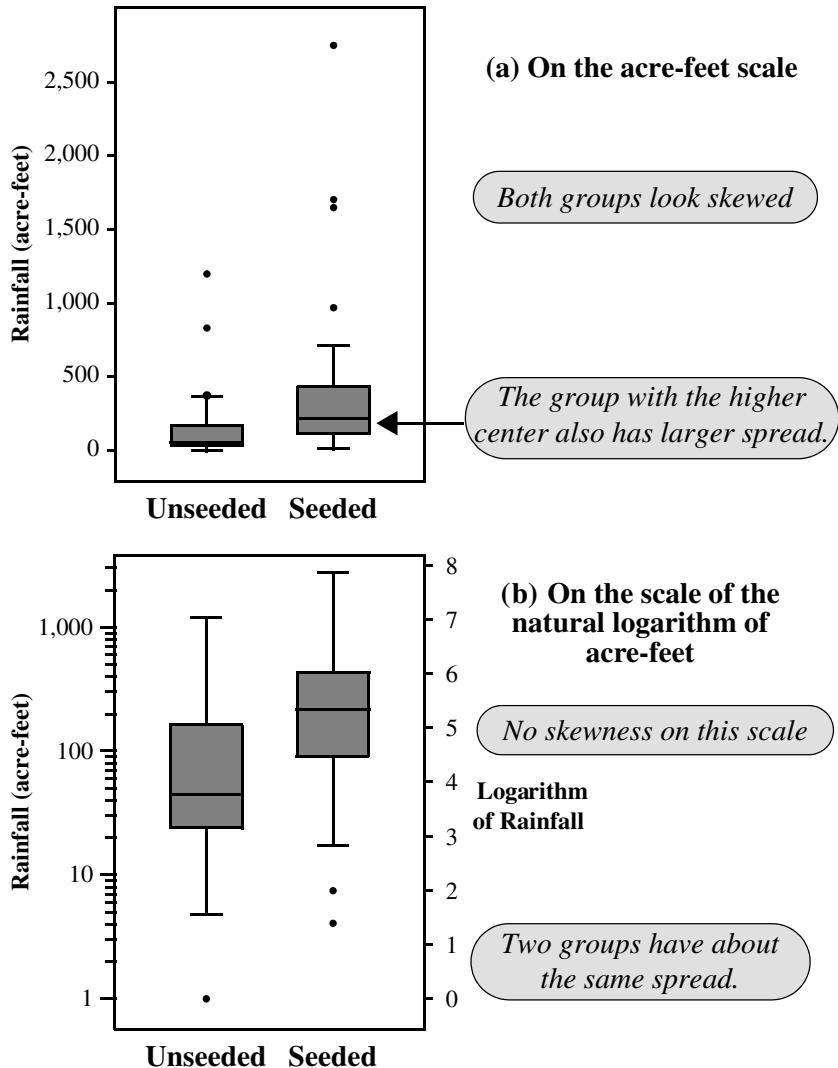
Box plots of the data in Display 3.2(a) indicate that the rainfall tended to be larger on the seeded days. Both distributions were quite skewed, and more variability occurred in the seeded group than in the control group. The box plots in Display 3.2(b) are drawn from the same data, but on the scale of the natural logarithm of the rainfall measurements. On this scale, the measurements appear to have symmetric distributions, and the variation seems nearly the same.

Statistical Conclusion

It is estimated that the volume of rainfall on days when clouds were seeded was 3.1 times as large as when not seeded. A 95% confidence interval for this multiplicative effect is 1.3 times to 7.7 times. Since randomization was used to determine whether any particular suitable day was seeded or not, it is safe to interpret this as evidence that the seeding caused the larger rainfall amount.

DISPLAY 3.2

Box plots of rainfall amounts on original and transformed scales



3.1.2 Effects of Agent Orange on Troops in Vietnam—An Observational Study

Many Vietnam veterans are concerned that their health may have been affected by exposure to Agent Orange, a herbicide sprayed in South Vietnam between 1962 and 1970. The particularly worrisome component of Agent Orange is a dioxin called TCDD, which in high doses is known to be associated with certain cancers. Studies have shown that high levels of this dioxin can be detected 20 or more years after

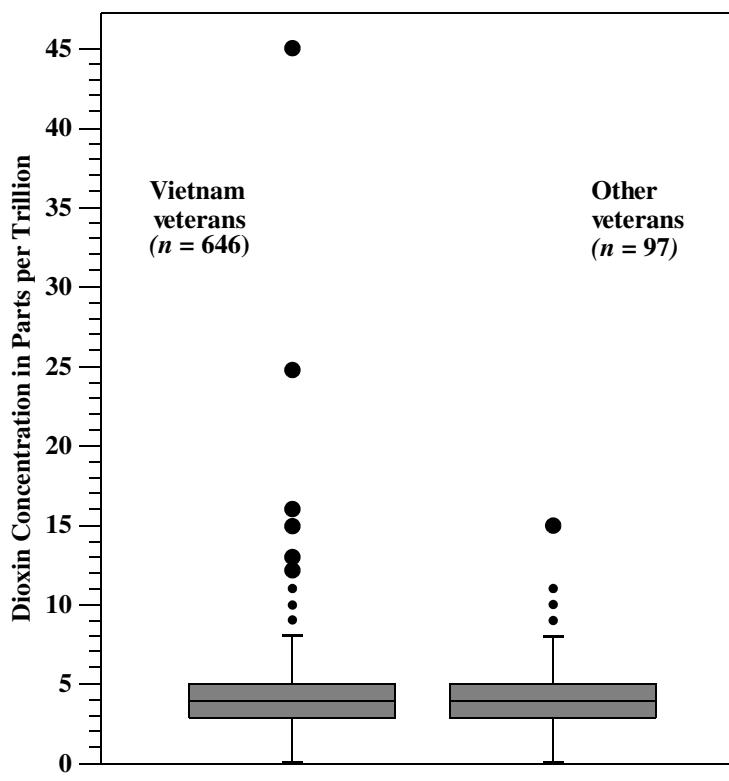
heavy exposure to Agent Orange. Consequently, as part of a series of studies, researchers from the Centers for Disease Control compared the current (1987) dioxin levels in living Vietnam veterans to the dioxin levels in veterans who did not serve in Vietnam.

The 646 Vietnam veterans in the study were a sample of U.S. Army combat personnel who served in Vietnam during 1967 and 1968, in the areas that were most heavily treated with Agent Orange. The 97 non-Vietnam veterans entered the Army between 1965 and 1971 and served only in the United States or Germany. Neither sample was randomly selected.

Blood samples from each veteran were analyzed for the presence of dioxin. Box plots of the observed levels are shown in Display 3.3. (Data from a graphical display in Centers for Disease Control Veterans Health Studies, "Serum 2,3,7,8-Tetrachlorodibenzo-*p*-dioxin Levels in U.S. Army Vietnam-era Veterans," *Journal of the American Medical Association* 260 (September 2, 1988): 1249–54.) The question of interest is whether the distribution of dioxin levels tends to be higher for the Vietnam veterans than for the non-Vietnam veterans.

DISPLAY 3.3

Box plots of 1987 dioxin concentrations in 646 Vietnam veterans and 97 veterans who did not serve in Vietnam



Statistical Conclusion

These data provide no evidence that the mean dioxin level in surviving Vietnam combat troops is greater than that for non–Vietnam veterans (one-sided p -value = 0.40, from a two-sample t -test). A 95% confidence interval for the difference in population means is –0.48 to 0.63 parts per trillion.

Scope of Inference

Since the samples were not random, inference to the populations is speculative. Participating veterans may not be representative of their respective groups. For example, nonparticipating Vietnam veterans may have failed to participate because of dioxin-related illnesses. If so, statistical statements about the populations of interest could be seriously biased. It should also be noted that many Vietnam veterans are frustrated and insulted by the prevalence of weak Agent Orange studies, like this one, which appear to address the Agent Orange problem but which actually skirt the main health issues.

3.2 ROBUSTNESS OF THE TWO-SAMPLE t -TOOLS

3.2.1 The Meaning of Robustness

The two-sample t -tools were used in the analyses of the Agent Orange study and the cloud seeding study, even though the actual conditions did not seem to match the ideal models upon which the tools are based. In the cloud seeding study, the t -tools were applied after taking the logarithms of the rainfalls. The t -tools could be used for the Agent Orange study, despite a lack of normality in the populations, because of the robustness of the t -tools against nonnormality.

*A statistical procedure is **robust to departures from a particular assumption** if it is valid even when the assumption is not met.*

Valid means that the uncertainty measures—the confidence levels and the p -values—are very nearly equal to the stated rates. For example, a procedure for obtaining a 95% confidence interval is valid if it is 95% successful in capturing the parameter. It is robust against nonnormality if it is roughly 95% successful with nonnormal populations.

Robustness of a tool must be evaluated separately for each assumption. The following sections detail the robustness of the two-sample t -tools against departures from the ideal assumptions of the normal, equal standard deviation model.

3.2.2 Robustness Against Departures from Normality

The *Central Limit Theorem* asserts that averages based on large samples have approximately normal sampling distributions, regardless of the shape of the population distribution. This suggests that underlying normality is not a serious issue, as long as sample sizes are reasonably large. The theorem provides only partial reassurance

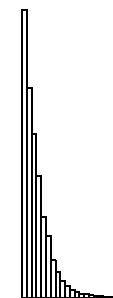
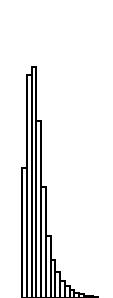
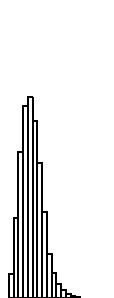
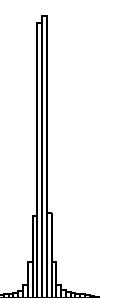
of applicability with respect to *t*-tools. It states what the sampling distribution of an average should be, but it does not address the effects of estimating a population standard deviation. Many empirical investigations and related theory, however, confirm that the *t*-tools remain reasonably valid in large samples, with many nonnormal populations.

How large is large enough? That depends on how nonnormal the population distributions are. Because distributions can differ from the normal in infinitely many ways, the question of sample size is difficult to answer. Statistical theory does say something fairly general about the relative effects of skewness and long-tailedness (*kurtosis*):

1. If the two populations have the same standard deviations and approximately the same shapes, and if the sample sizes are about equal, then the validity of the *t*-tools is affected moderately by long-tailedness and very little by skewness.
2. If the two populations have the same standard deviations and approximately the same shapes, but if the sample sizes are not approximately the same, then the validity of the *t*-tools is affected moderately by long-tailedness and substantially by skewness. The adverse effects diminish, however, with increasingly large sample sizes.
3. If the skewness in the two populations differs considerably, the tools can be very misleading with small and moderate sample sizes.

Computer simulations can clarify the role of sample sizes. To further investigate the effect of nonnormality on 95% confidence intervals, a computer was instructed to generate samples from the nonnormal distributions shown in Display 3.4. For each pair of generated samples, it computed the 95% confidence interval for the difference in population means and recorded whether the interval actually captured the *true* difference in population means. The actual percentage of successful intervals from 1,000 simulations is shown in Display 3.4 for each set of conditions examined. The purpose of the simulation is to identify combinations of sample sizes and nonnormally shaped distributions for which the confidence interval procedure has a success rate of nearly 95%. An actual success rate less than 95% is undesirable because it means the intervals tend to be too short for the given confidence level and they therefore tend to exclude possible parameter values that shouldn't be excluded. The corresponding hypothesis test, in this case, tends to produce more false claims about statistical significance than expected. An actual success rate greater than 95% is also undesirable because it means the intervals tend to be too long and include possible parameter values that shouldn't be included. The corresponding hypothesis test, in this case, tends to miss statistically significant findings that it really should find.

Of the five distributions examined, only the long-tailed distribution appears to have success rates that are poor enough to cause potentially misleading statements—and even those are not too bad. This distribution can be recognized in practice by the presence of outliers. For the skewed distributions, however, the normality assumption does not appear to be a major concern even for small sample sizes, at least as long as the skewness is the same in the two populations and the sample sizes are roughly equal.

		Percentage of 95% confidence intervals that are successful when the two populations are non-normal (but with same shape and SD, and equal sample sizes) (each percentage is based on 1,000 computer simulations)				
		Strongly skewed	Moderately skewed	Mildly skewed	Long-tailed	Short-tailed
Sample size						
5	95.5	95.4	95.2	98.3	94.5	
10	95.5	95.4	95.2	98.3	94.6	
25	95.3	95.3	95.1	98.2	94.9	
50	95.1	95.3	95.1	98.1	95.2	
100	94.8	95.3	95.0	98.0	95.6	

3.2.3 Robustness Against Differing Standard Deviations

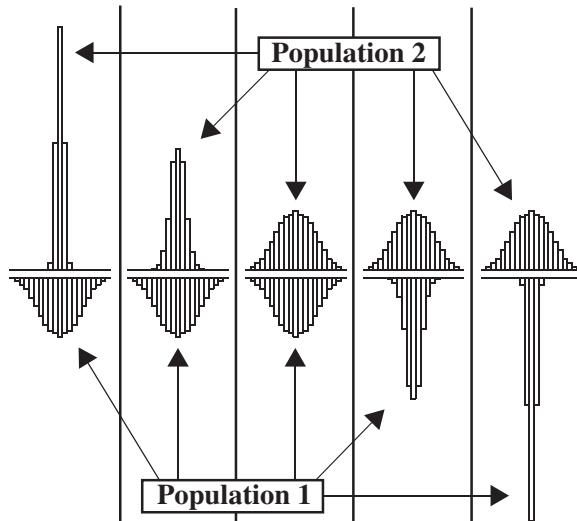
More serious problems may arise when the standard deviations of the two populations are substantially unequal. In this case, the pooled estimate of standard deviation does not estimate any population parameter and the standard error formula, which uses the pooled estimate of standard deviation, no longer estimates the standard deviation of the difference between sample averages. As a result, the *t*-ratio does not have a *t*-distribution.

Theory shows that the *t*-tools remain fairly valid when the standard deviations are unequal, as long as the sample sizes are roughly the same. For clarification, a computer was again instructed to generate pairs of samples, this time from two normal populations with different standard deviations, as shown in Display 3.5. It computed 95% confidence intervals for each pair of samples and recorded whether the resulting interval successfully captured the true difference in population means. The actual percentages successful are displayed.

Notice that the success rates for the rows with equal sample sizes ($n_1 = n_2 = 10$ and $n_1 = n_2 = 100$) are very nearly 95%. Thus, as suggested by theory, unequal population standard deviations have little effect on validity if the sample sizes are equal. For substantially different σ 's and different n 's, however, the confidence intervals are unreliable. The worst situation is when the ratio of standard deviations is much different from 1 and the smaller sized sample is from the population with the larger standard deviation (as, for example, when $n_1 = 100$, $n_2 = 400$, and $\sigma_2/\sigma_1 = 1/4$).

DISPLAY 3.5

Percentage of successful 95% confidence intervals when the two populations have different standard deviations (but are normal) with possibly different sample sizes (each percentage is based on 1,000 computer simulations)



n_1	n_2	$\sigma_2/\sigma_1=1/4$	$\sigma_2/\sigma_1=1/2$	$\sigma_2/\sigma_1=1$	$\sigma_2/\sigma_1=2$	$\sigma_2/\sigma_1=4$
10	10	95.2	94.2	94.7	95.2	94.5
10	20	Success rates for 95% intervals	83.0	89.3	94.4	98.7
10	40		71.0	82.6	95.2	99.5
100	100		94.8	96.2	95.4	95.3
100	200		86.5	88.3	94.8	98.8
100	400		71.6	81.5	95.0	99.5

3.2.4 Robustness Against Departures from Independence

Cluster Effects and Serial Effects

Whenever knowledge that one observation is, say, above average allows an improved guess about whether another observation will be above average, *independence* is lacking. The methods of Chapter 2 may be misleading in this case. Two types of dependence (lack of independence) commonly arise in practical problems.

The first is a *cluster effect*, which sometimes occurs when the data have been collected in subgroups. For example, 50 experimental animals may have been collected from 10 litters and then randomly assigned to one of two treatment groups. Since animals from the same litter may tend to be more similar in their responses than animals from different litters, it is likely that independence is lacking.

The other type of dependence commonly encountered is caused by a *serial effect*, in which measurements are taken over time and observations close together in time tend to be more similar (or perhaps more different) than observations collected

at distant time points. This can also occur if measurements are made at different locations, and measurements physically close to each other tend to be more similar than those farther apart. In the latter case the dependence pattern is called *spatial correlation*.

The Effects of Lack of Independence on the Validity of the *t*-Tools

When the independence assumptions are violated, the standard error of the difference of averages is an inappropriate estimate of the standard deviation of the difference in averages. The *t*-ratio no longer has a *t*-distribution, and the *t*-tools may give misleading results. The seriousness of the consequences depends on the seriousness of the violation. It is generally unwise to use the *t*-tools directly if cluster or serial effects are suspected. Other methods that adjust for these effects are available (Chapters 9–15).

3.3 RESISTANCE OF THE TWO-SAMPLE *t*-TOOLS

Some practical suggestions will soon be provided for sizing up the actual conditions and choosing a course of action. The effect of outliers on the *t*-tools is discussed first, however, since decisions about how to deal with the outliers play an important role in the overall strategy.

3.3.1 Outliers and Resistance

An *outlier* is an observation judged to be far from its group average. The effect of outliers on the two-sample *t*-tools has partially been addressed in the discussion of robustness. In fact, it is evident from the theoretical results and the computer simulations in Display 3.4 that the *p*-values and confidence intervals may be unreliable if the population distributions are long-tailed. Since long-tailed distributions are characterized by the presence of outliers in the sample, outliers should cause some concern.

Long-tailed population distributions are not the only explanation for outliers, however. The populations of interest may be normal but the sample may be contaminated by one or more observations that do not come from the population of interest. Often it is philosophically difficult and practically irrelevant to distinguish between a natural long-tailed distribution and one that includes outliers that result from contamination, although in some cases the identification of clear contamination may dictate an obvious course of action. For example, if it is discovered that one member of a sample from a population of 25- to 35-year-old women is, in fact, over 50 years old, she should be removed from the sample.

It is useful to know how sensitive a statistical procedure may be to one or two outlying observations. The notion of resistance addresses this issue:

*A statistical procedure is **resistant** if it does not change very much when a small part of the data changes, perhaps drastically.*

As an example, consider the hypothetical sample: 10, 20, 30, 50, 70. The sample average is 36, and the sample median is 30. Now change the 70 to 700, and what happens? The sample average becomes 162, but the sample median remains 30. The sample average is not a resistant statistic because it can be severely influenced by the change in a single observation. The median, however, is resistant.

Resistance is a desirable property. A resistant procedure is insensitive to outliers. A nonresistant one, on the other hand, may be greatly influenced by one or two outlying observations.

3.3.2 Resistance of *t*-Tools

Since *t*-tools are based on averages, they are not resistant. A small portion of the data can potentially have a major influence on the results. In particular, one or two outliers can affect a confidence interval or change a *p*-value enough to completely alter a conclusion.

If the outlier is due to contamination from another population, it can lead to false impressions about the population of interest. If the outlier does come from the population of interest, which happens to be long-tailed, the outcome is still undesirable for the following reason. In statistics, the goal is to describe *group* characteristics. An estimate of the center of a distribution should represent the typical value. The estimate is a good one if it represents the typical values possessed by the great majority of subjects; it is a bad one if it represents a feature unique to one or two subjects. Furthermore, a conclusion that hinges on one or two data points must be viewed as quite fragile.

3.4 PRACTICAL STRATEGIES FOR THE TWO-SAMPLE PROBLEM

Armed with information about the broad set of conditions under which the *t*-tools work well and the effect of outliers, the challenge to the data analyst is to size up the actual conditions using the available data and evaluate the appropriateness of the *t*-tools. This involves thinking about possible cluster and serial effects; evaluating the suitability of the *t*-tools by examining graphical displays; and considering alternatives.

In considering alternatives it is important to realize that even though the *t*-tools may still be valid when the ideal assumptions are not met, an alternative procedure that is more *efficient* (i.e., makes better use of the data) may be available. For example, another procedure may provide a narrower confidence interval.

Consider Serial and Cluster Effects

To detect lack of independence, carefully review the method by which the data were gathered. Were the subjects selected in distinct groups? Were different groups of subjects treated differently in a way that was unrelated to the primary treatment? Were different responses merely repeated measurements on the same subjects? Were observations taken at different but proximate times or locations? Affirmative answers to any of these questions suggest that independence may be lacking.

The principal remedy is to use a more sophisticated statistical tool. Identifiable clusters, which may be planned or unplanned, can be accounted for through analysis

of variance (Chapters 13 and 14) or possibly through regression analysis (Chapters 9–12). Serial effects require time series analysis, the topic of Chapter 15.

Evaluate the Suitability of the *t*-Tools

Side-by-side histograms or box plots of the two groups of data should be examined and departures from the ideal model should be considered in light of the robustness properties of the *t*-tools. It is important to realize that the conditions of interest, which are those of the populations, must be investigated through graphical displays of the samples.

If the conditions do not appear suitable for use of the *t*-tools, then some alternative is necessary. A transformation should be considered if the graphical displays of the transformed data appear to be closer to the ideal conditions. (See Section 3.5.) Alternative tools for analyzing two independent samples are the rank-sum procedure, which is resistant and does not depend on normality (Section 4.2); other permutation tests (Section 4.3.1); and the Welch procedure for comparing normal populations that have unequal standard deviations (Section 4.3.2).

A Strategy for Dealing with Outliers

If investigation reveals that an outlying observation was recorded improperly or was the result of contamination from another population, the solution is to correct it if the right value is known or to leave it out. Often, however, there is no way to know how the outliers arose. Two statistical approaches for dealing with this situation exist. One is to employ a resistant statistical tool, in which case there is no compelling reason to ponder whether the offending observations are natural, the result of contamination, or simply blunders. (The rank-sum procedure in Section 4.2 is resistant.) The other approach is to adopt the careful examination strategy shown in Display 3.6. An important aspect of adopting this procedure is that an outlier does not get swept under the rug simply because it is different from the other observations. To warrant its removal, an explanation for why it is different must be established.

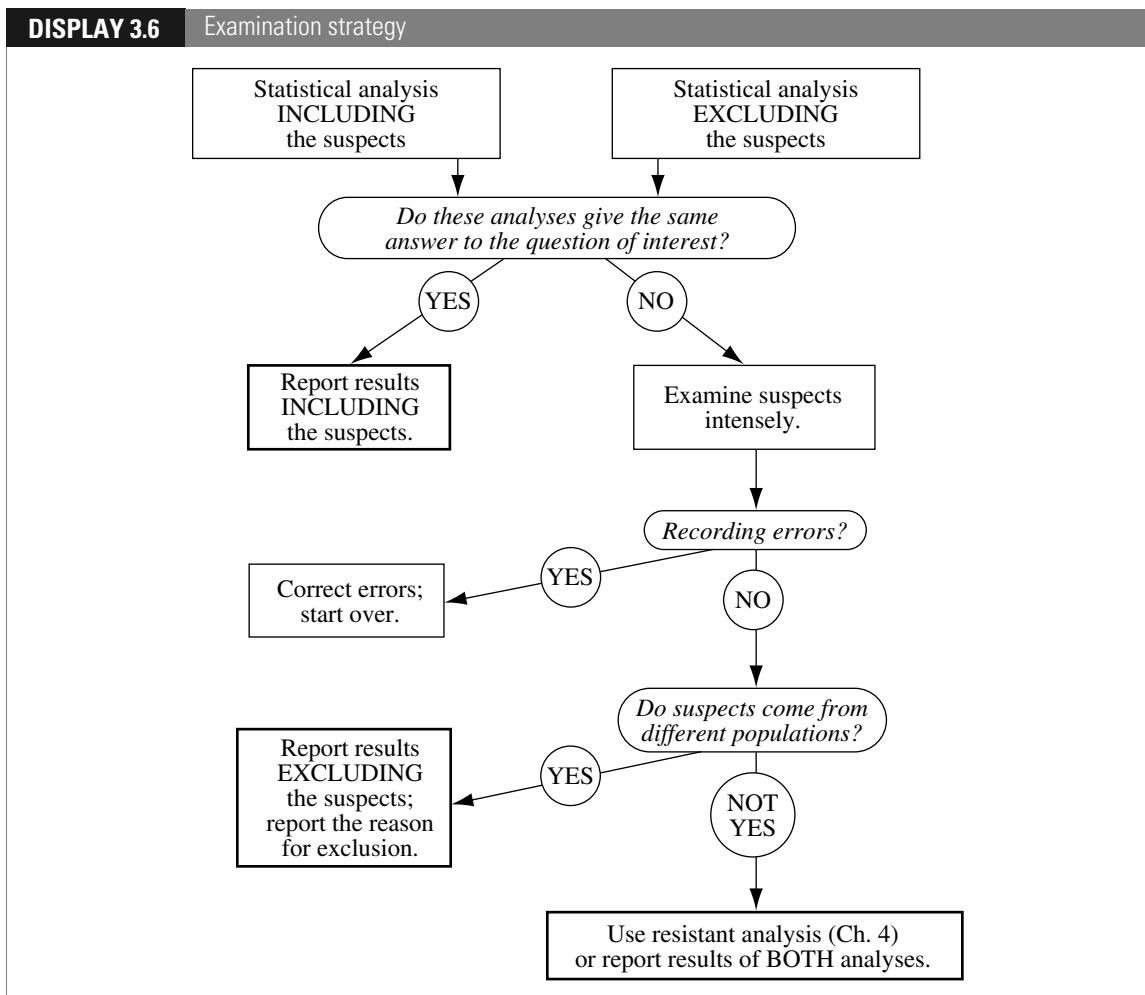
Example—Agent Orange

Box plots of dioxin levels in Vietnam and non–Vietnam veterans (Display 3.3) appear again in Display 3.7. The distributions have about the same shape and spread. Although the shape is not normal, the skewness is mild and unlikely to cause any problems with the *t*-test or the confidence interval. Two Vietnam veterans (#645 and #646) had considerably higher dioxin levels than the others.

From the results listed in Display 3.7 it is evident that the comparison of the two groups is changed very little by the removal of one or both of these outliers. Consequently, there is no need for further action. Even so, it is useful to see what else can be learned about these two, as indicated at the bottom of the display.

Notes

1. It is not useful to give a precise definition for an *outlier*. Subjective examination is the best policy. If there is any doubt about whether a particular observation deserves further examination, give it further examination.



2. It is not surprising that the outliers in the Agent Orange example have little effect, since the sample sizes are so large.
3. The apparent difference in the box plots may be due to the difference in sample sizes. If the population distributions are identical, more observations will appear in the extreme tails from a sample of size 646 than from a sample of size 97.

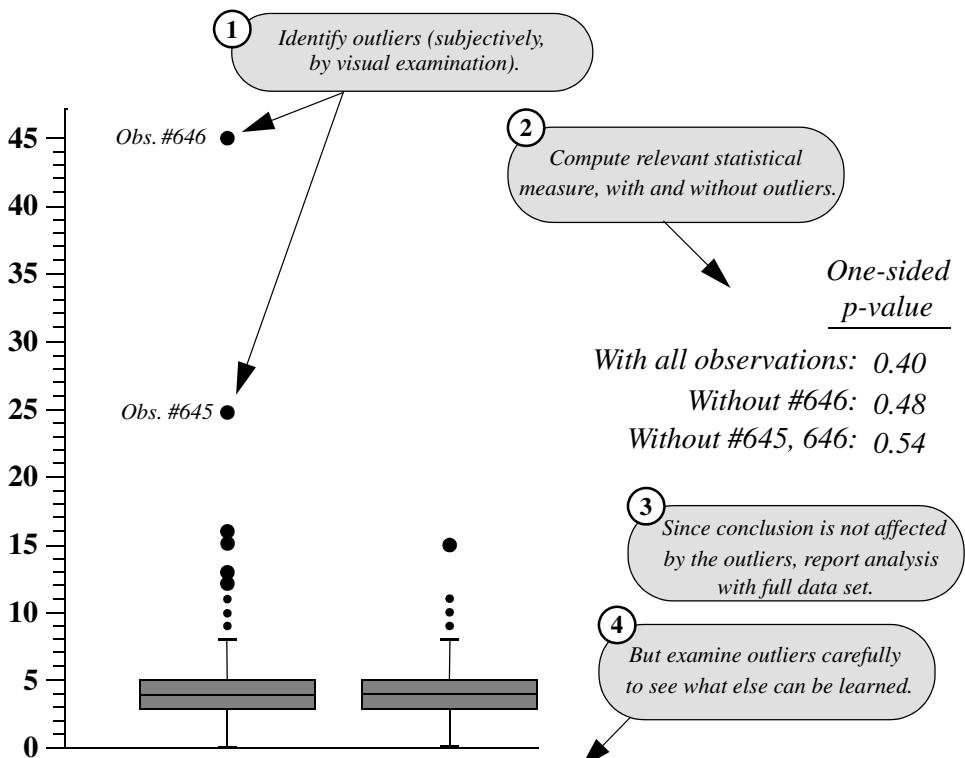
3.5 TRANSFORMATIONS OF THE DATA

3.5.1 The Logarithmic Transformation

The most useful transformation is the *logarithm* (log) for positive data. The common scale for scientific work is the *natural logarithm* (ln), based on the number

DISPLAY 3.7

Outlier analysis for Agent Orange data: effect of outliers on the p -value, for equal population means



Veteran # 645: reported 180 days of indirect military exposure to herbicides.

Veteran # 646: reported no exposure (military or civilian) to herbicides.

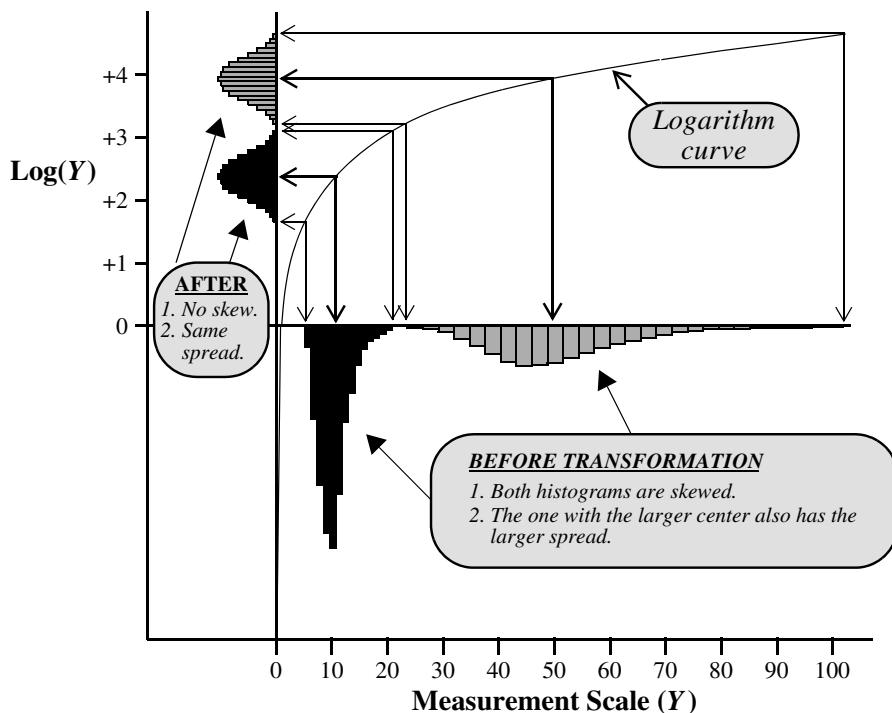
$e = 2.71828\dots$. The logarithm of e is unity, denoted by $\log(e) = 1$. Also, the log of 1 is 0: $\log(1) = 0$. The general rule for using logarithms is that $\log(e^x) = x$. Another choice is the *common* logarithm based on the number 10, rather than e . Common logs are defined by $\log_{10}(10^x) = x$. Unless otherwise stated, *log* in this book refers to the natural logarithm.

Recognizing the Need for a Log Transformation

The data themselves usually suggest the need for a log transformation. If the ratio of the largest to the smallest measurement in a group is greater than 10, then the data are probably more conveniently expressed on the log scale. Also, if the graphical displays of the two samples show them both to be skewed and if the group with the larger average also has the larger spread (see Display 3.2), the log transformation is likely to be a good choice.

DISPLAY 3.8

The logarithmic transformation used to arrive at favorable conditions for the two-sample t -analysis



Display 3.8 illustrates the behavior of the log transformation. On the scale of measurement Y the two groups have skewed distributions with longer tails in the positive direction. The group with the larger center also has the larger spread. The measurements on the transformed scale have the same ordering, but small numbers get spread out more, while large numbers are squeezed more closely together. The overall result is that the two distributions on the transformed scale appear to be symmetric and have equal spread—just the right conditions for applying the t -tools.

3.5.2 Interpretation After a Log Transformation

For some measurements, the results of an analysis are appropriately presented on the transformed scale. Most users feel comfortable with the Richter scale for measuring earthquake strength, even though it is a logarithmic scale. Similarly, pH as a measure of acidity is the negative log of ion concentration. In other cases, however, it may be desirable to present the results on the original scale of measurement.

Randomized Experiment Model: Multiplicative Treatment Effect

If the randomized experiment model with additive treatment effect is thought to hold for the log-transformed data, then an experimental unit that would respond

to treatment 1 with a logged outcome of $\log(Y)$ would respond to treatment 2 with a logged outcome of $\log(Y) + \delta$. By taking antilogarithms of these two quantities, one finds that an experimental unit that would respond to treatment 1 with an outcome of Y would respond to treatment 2 with an outcome of Ye^δ . Thus, e^δ is the *multiplicative treatment effect* on the original scale of measurement. To test whether there is any treatment effect, one performs the usual t -test for the hypothesis that δ is zero with the log-transformed data. To describe the multiplicative treatment effect, one back-transforms the estimate of δ and the endpoints of the confidence interval for δ .

Interpretation After Log Transformation (Randomized Experiment)

Suppose $Z = \log(Y)$. It is estimated that the response of an experimental unit to treatment 2 will be $\exp(\bar{Z}_2 - \bar{Z}_1)$ times as large as its response to treatment 1.

Example—Cloud Seeding

Display 3.2 shows that the log-transformed rainfalls have distributions that appear satisfactory for using the t -tools; so in Display 3.9 a full analysis is carried out on the log scale. Tests and confidence intervals are constructed in the usual way but on the transformed data. The estimate of the additive treatment effect on log rainfall is back-transformed to an estimate of the multiplicative effect of cloud seeding on rainfall.

Population Model: Estimating the Ratio of Population Medians

The t -tools applied to log-transformed data provide inferences about the difference in means of the logged measurements, which may be represented as $\text{Mean}[\log(Y_2)] - \text{Mean}[\log(Y_1)]$, where $\text{Mean}[\log(Y_2)]$ symbolizes the mean of the logged values of population 2. A problem with interpretation on the original scale arises because the mean of the logged values is not the log of the mean. Taking the antilogarithm of the estimate of the mean on the log scale does *not* give an estimate of the mean on the original scale.

If, however, the log-transformed data have symmetric distributions, the following relationships hold:

$$\text{Mean}[\log(Y)] = \text{Median}[\log(Y)]$$

(and since the log preserves ordering)

$$\text{Median}[\log(Y)] = \log[\text{Median}(Y)],$$

where $\text{Median}(Y)$ represents the *population median* (the 50th percentile of the population). In other words, the 50th percentile of the logged values is the log of the 50th percentile of the untransformed values. Putting these two equalities together,

DISPLAY 3.9

Two-sample t -analysis and statement of conclusions after logarithmic transformation—cloud seeding example

1 Transform the data.

Unseeded		Seeded	
Y (acre-ft)	$\log(Y)$	Y (acre-ft)	$\log(Y)$
1202.6	7.092	2745.6	7.918
830.1	6.722	1697.8	7.437
372.4	5.920	1656.0	7.412
345.5	5.845	978.0	6.886
321.2	5.772	703.4	6.556
244.3	5.498	489.1	6.193
163.0	5.094	430.0	6.064
147.8	4.996	334.1	5.811
95.0	4.554	302.8	5.713
87.0	4.466	274.7	5.616
81.2	4.397	274.7	5.616
68.5	4.227	255.0	5.541
47.3	3.857	242.5	5.491
41.1	3.716	200.7	5.302
36.6	3.600	198.6	5.291
29.0	3.367	129.6	4.864
28.6	3.353	119.0	4.779
26.3	3.270	118.3	4.773
26.1	3.262	115.3	4.748
24.4	3.195	92.4	4.526
21.7	3.077	40.6	3.704
17.3	2.851	32.7	3.487
11.5	2.446	31.4	3.447
4.9	1.589	17.5	2.862
4.9	1.589	7.7	2.041
1.0	0.000	4.1	1.411

2 Use the two-sample t -tools on the log rainfall.

Difference in averages = 1.1436 (SE = 0.4495).

Test of the hypothesis of no effect of cloud seeding on log rainfall: one-sided p -value from two-sample t -test = 0.0070 (50 d.f.).

95% confidence interval for additive effect of cloud seeding on log rainfall: 0.2406 to 2.0467.

3 Back-transform estimate and confidence interval.

$$\text{Estimate} = e^{1.1436} = 3.1382$$

$$\text{Lower confidence limit} = e^{0.2406} = 1.2720.$$

$$\text{Upper confidence limit} = e^{2.0467} = 7.7425.$$

4 State the conclusions on the original scale.

Conclusion: There is convincing evidence that seeding increased rainfall (one-sided p -value = 0.0070). The volume of rainfall produced by a seeded cloud is estimated to be 3.14 times as large as the volume that would have been produced in the absence of seeding (95% confidence: 1.27 to 7.74 times).

it is evident that the antilogarithm of the mean of the log values is the median on the original scale of measurements.

If \bar{Z}_1 and \bar{Z}_2 are used to represent the averages of the logged values for samples 1 and 2, then $\bar{Z}_2 - \bar{Z}_1$ estimates $\log[\text{Median}(Y_2)] - \log[\text{Median}(Y_1)]$, and therefore

$$\bar{Z}_2 - \bar{Z}_1 \text{ estimates } \log \left[\frac{\text{Median}(Y_2)}{\text{Median}(Y_1)} \right]$$

and, therefore,

$$\exp(\bar{Z}_2 - \bar{Z}_1) \text{ estimates } \left[\frac{\text{Median}(Y_2)}{\text{Median}(Y_1)} \right].$$

The point of this is that a very useful multiplicative interpretation emerges in terms of the ratio of population medians. This is doubly important because the median is a better measure of the center of a skewed distribution than the mean. The multiplicative nature of this relationship is captured with the following wording:

**Interpretation After Log Transformation
(Observational Study)**

It is estimated that the median for population 2 is $\exp(\bar{Z}_2 - \bar{Z}_1)$ times as large as the median for population 1.

In addition, back-transforming the ends of a confidence interval constructed on the log scale produces a confidence interval for the ratio of medians.

Example (Sex Discrimination)

Although the analysis of the sex discrimination data of Section 1.1.2, was suitable on the original scale of the untransformed salaries, graphical displays of the log-transformed salaries indicate that analysis would also be suitable on the log scale. The average male log salary minus the average female log salary is 0.147. Since $e^{0.147} = 1.16$, it is estimated that the median salary for males is 1.16 times as large as the median salary for females. Equivalently, the median salary for males is estimated to be 16% more than the median salary for females. Since a 95% confidence interval for the difference in means on the log scale is 0.100 to 0.194, a 95% confidence interval for the ratio of population median salaries is 1.11 to 1.21 ($e^{0.100}$ to $e^{0.194}$). With 95% confidence, it is estimated that the median salary for males is between 11% and 21% greater than the median salary for females.

3.5.3 Other Transformations for Positive Measurements

There are other useful transformations for positive measurements with skewed distributions where the means and standard deviations differ between groups. The *square root* transformation \sqrt{Y} applies to data that are counts—counts of bacteria clusters in a dish, counts of traffic accidents on a stretch of highway, counts of red giants in a region of space—and to data that are measurements of area. The *reciprocal* transformation $1/Y$ applies to data that are waiting times—times to failure of lightbulbs, times to recurrence for cancer patients treated with radiation, reaction times to visual stimuli, and so on. The reciprocal of a time measurement can often be interpreted directly as a rate or a speed. The *arcsine square root* transformation, $\text{arcsine}(\sqrt{Y})$, and the *logit* transformation, $\log[Y/(1 - Y)]$, apply when the measurements are proportions between zero and one—proportions of trees infested by

a wood-boring insect in experimental plots, proportions of weight lost as a side effect of leukemia therapy, proportions of winning lottery tickets in clusters of a certain size, and so forth.

Only the log transformation, however, gives such ease in converting inferences back to the original scale of measurement. One may estimate the difference in means of $\sqrt{Y_2}$ and $\sqrt{Y_1}$, but the square of this difference does not make much sense on the original scale.

Choosing a Transformation

Formal statistical methods are available for selecting a transformation. Nevertheless, it is recommended here that a trial-and-error approach, with graphical analysis, be used instead. For positive data in need of a transformation, the logarithm should almost always be the first tried. If it is not satisfactory, the reciprocal or the square root transformations might be useful. Keep in mind that the primary goal is to establish a scale where the two groups have roughly the same spread. If several transformations are similar in their ability to accomplish this, think carefully about which one offers the most convenient interpretation.

Caveat About the Log Transformation

Situations arise where presenting results in terms of population medians is not sufficient. For example, the daily emissions of dioxin in the effluent from a paper mill have a very skewed distribution. An agency monitoring the emissions will be interested in estimating the total dioxin load released during, say, a year of operation. The total dioxin load would be the population mean times the population size, and therefore is estimated by the sample average times the population size. It cannot be estimated directly from the median, unless more specific assumptions are made.

3.6 RELATED ISSUES

3.6.1 Prefer Graphical Methods Over Formal Tests for Model Adequacy

Formal tests for judging the adequacy of various assumptions exist. Tests for normality and tests for equal standard deviation are available in most statistical computer programs, as are tests that determine whether an observation is an outlier. Despite their widespread availability and ease of use, these diagnostic tests are not very helpful for model checking. They reveal little about whether the data meet the broader conditions under which the tools work well. The fact that two populations are not exactly normal, for example, is irrelevant. Furthermore, the formal tests themselves are often not very robust against their own model assumptions. Graphical displays are more informative, if less formal. They provide a good indication of whether or not the data are amenable to t -analysis and, if not, they often suggest a remedy.

3.6.2 Robustness and Transformation for Paired *t*-Tools

The one-sample *t*-test, of which the paired *t*-test is a special case, assumes that the observations are independent of one another and come from a normally distributed population. *P*-values and confidence intervals remain valid for moderate and large sample sizes for nonnormal distributions. For smaller sample sizes skewness can be a problem. When cluster or serial effects are present (see Section 3.2.4), the *t*-tools may give misleading results. When the observations within each pair are positive, either an apparent multiplicative treatment effect (in an experiment) or a tendency for larger differences in pairs with larger average values suggests the use of a log transformation. The transformation is applied before taking the difference, which is equivalent to forming a ratio within each pair and performing a one-sample analysis on the logarithms of the ratios. If there are n pairs, let $Z_i = \log(Y_{1i}) - \log(Y_{2i})$, which is the same as $\log(Y_{1i}/Y_{2i})$. In an observational study, $\exp(\bar{Z})$ is an estimate of the median of the ratios, Y_1/Y_2 . (This is not the same as the ratio of the medians [see Exercise 20].) In a randomized, paired experiment, $\exp(\bar{Z})$ estimates a multiplicative treatment effect on the original scale. In both cases, the statistical work of testing and constructing a confidence interval is done on the log scale. The estimate and associated interval are transformed back to the original scale.

3.6.3 Example—Schizophrenia

In the schizophrenia example of Section 2.1.2, Z_i represents the logarithm of the left hippocampus volume of the unaffected twin divided by the left hippocampus volume of the affected twin in pair i . The average of the 15 log ratios is 0.1285. A one-sample analysis gives a *p*-value of 0.0065 for the test that the mean is zero and a 95% confidence interval from 0.0423 to 0.2147 for the mean itself. Taking antilogarithms of the estimate and the endpoints of the confidence interval yields the following conclusion: It is estimated that the median of the unaffected-to-affected volume ratios is 1.137. A 95% confidence interval for the median ratio is from 1.043 to 1.239.

3.7 SUMMARY

Cloud Seeding and Rainfall Study

The box plots of the rainfalls for seeded and unseeded days reveal that the two distributions of rainfall are skewed and that the distribution with the larger mean also has the larger variance. This is the situation where log-transformed data behave in accordance with the ideal model. A plot of the data after transformation confirms the adequacy of the transformation. The two-sample *t*-test can be used as an approximation to the randomization test, and the difference in averages (of log rainfall) can be back-transformed to provide a statement about a multiplicative treatment effect. In the example, it is estimated that the rainfall is 3.1 times as much when a cloud is seeded as when it is left unseeded.

Since randomization is used, the statistical conclusion implies that the seeding causes the increase in rainfall. Since the decision about whether to seed clouds is determined (in this case) by a random mechanism, and since the airplane crew is *blind* to which treatment they are administering, human bias can have had little influence on the result.

Agent Orange Study

Graphical analysis focuses attention on the possibly undue influence of two outliers, but analyses with and without the outliers reveal no such influence, so the *t*-tools are used on the entire data set. The form of the sampling from the populations of living Vietnam veterans and of other veterans is a major concern in accepting the reliability of the statistical analysis. Protocols for obtaining the samples have not been discussed here, except to note that random sampling is not being used. Conclusions based on the two-sample *t*-test are supplied, along with the caveat that there may be biases due to the lack of random sampling.

3.8 EXERCISES

Conceptual Exercises

1. **Cloud Seeding.** What is the experimental unit in the cloud seeding experiment?
2. **Cloud Seeding.** Randomization in the cloud seeding experiment was crucial in assessing the effect of cloud seeding on rainfall. Why?
3. **Cloud Seeding.** Why was it important that the airplane crew was unaware of whether seeding was conducted or not?
4. **Cloud Seeding.** Why would it be helpful to have the date of each observed rainfall?
5. **Agent Orange.** How would you respond to the comment that the box plots in Display 3.6 indicate that the dioxin levels in the Vietnam veterans tend to be larger since their values appear to be larger?
6. **Agent Orange.** (a) What course of action would you propose for the statistical analysis if it was learned that Vietnam veteran #646 (the largest observation in Display 3.6) worked for several years, after Vietnam, handling herbicides with dioxin? (b) What would you propose if this was learned instead for Vietnam veteran #645?
7. **Agent Orange.** If the statistical analysis had shown convincing evidence that the mean dioxin levels differed in Vietnam veterans and other veterans, could one conclude that serving in Vietnam was responsible for the difference?
8. **Schizophrenia.** In the schizophrenia study in Section 2.1.2, the observations in the two groups (schizophrenic and nonschizophrenic) are not independent since each subject is matched with a twin in the other group. Did the researchers make a mistake?
9. True or false? A statistical computer package will only print out a *p*-value or confidence interval if the conditions for its validity are met.
10. True or false? A sample histogram will have a normal distribution if the sample size is large enough.

- 11.** A woman who has just moved to a new job in a new town discovers two routes to drive from her home to work. The first Monday, she flips a coin, deciding to take route A if it comes up heads and to take route B if it is tails. The following Monday, she will take the other route. The first Tuesday, she flips the coin again with the same plan. And so on for the first week. At the end of two weeks, she has traveled both routes five times and can compare their average commuting times. Why should she not use the *t*-tools for two independent samples? What should she use?
- 12.** In which ways are the *t*-tools more robust for larger sample sizes than for smaller ones (i.e., robust with respect to normality, equal SDs, and/or independence)?
- 13. Fish Oil.** Why is a log transformation inappropriate for the fish oil data in Exercise 1.12?
- 14.** Will an outlier from a contaminating population be more consequential in small samples or large samples?
- 15.** What would you suggest as an alternative estimate of the standard deviation of the difference in sample averages when it is clear that the two populations have different SDs? (Check the formula for the standard deviation of the sampling distribution of the difference in averages, in Display 2.6.)
- 16.** A researcher has taken tissue cultures from 25 subjects. Each culture is divided in half, and a treatment is applied to one of the halves chosen at random. The other half is used as a control. After determining the percent change in the sizes of all culture sections, the researcher calculates the standard error for the treatment-minus-control difference using both the paired *t*-analysis and the two independent sample (Chapter 2) *t*-analysis. Finding that the paired *t*-analysis gives a slightly larger standard error (and gives only half the degrees of freedom), the researcher decides to use the results from the unpaired analysis. Is this legitimate?
- 17.** Respiratory breathing capacity of individuals in houses with low levels of nitrogen dioxide was compared to the capacity of individuals in houses with high levels of nitrogen dioxide. From a sample of 200 houses of each type, breathing capacity was measured on 600 individuals from houses with low nitrogen dioxide and on 800 individuals from houses with high nitrogen dioxide. (a) What problem do you foresee in applying *t*-tools to these data? (b) Would comparing the average *household* breathing capacities avoid the problem?
- 18. Trauma and Metabolic Expenditure.** The following data are metabolic expenditures for eight patients admitted to a hospital for reasons other than trauma and for seven patients admitted for multiple fractures (trauma). (Data from C. L. Long, et al., "Contribution of Skeletal Muscle Protein in Elevated Rates of Whole Body Protein Catabolism in Trauma Patients," *American Journal of Clinical Nutrition* 34 (1981): 1087–93.)

Metabolic Expenditures (kcal/kg/day)

Nontrauma patients:	20.1	22.9	18.8	20.9	20.9	22.7	21.4	20.0
Trauma patients:	38.5	25.8	22.0	23.0	37.6	30.0	24.5	

- (a) Is the difference in averages resistant? (*Hint:* What happens if 20.0 is replaced by 200?)
 (b) Replacing each value with its rank, from the lowest to highest, in the combined sample gives

Metabolic Expenditures (kcal/kg/day)

Nontrauma patients:	3	9	1	4.5	4.5	8	6	2
Trauma patients:	15	12	7	10	14	13	11	

Consider the average of the ranks for the trauma group minus the average of the ranks for the nontrauma group. Is this statistic resistant?

19. In each of the following data problems there is some potential violation of one of the independence assumptions. State whether there is a cluster effect or serial correlation, and whether the questionable assumption is the independence within groups or the independence between groups.

- (a) Researchers interested in learning the effects of speed limits on traffic accidents recorded the number of accidents per year for each of 10 consecutive years on roads in a state with speed limits of 90 km/h. They also recorded the number of accidents for the next 7 years on the same roads after the speed limit had been increased to 110 km/hr. The two groups of measurements are the number of accidents per year for those years under study. (Notice that there is also a potential confounding variable here!)
- (b) Researchers collected intelligence test scores on twins, one of whom was raised by the natural parents and one of whom was raised by foster parents. The data set consists of test scores for the two groups, boys raised by their natural parents and boys raised by foster parents.
- (c) Researchers interested in investigating the effect of indoor pollution on respiratory health randomly select houses in a particular city. Each house is monitored for nitrogen dioxide concentration and categorized as being either high or low on the nitrogen dioxide scale. Each member of the household is measured for respiratory health in terms of breathing capacity. The data set consists of these measures of respiratory health for all individuals from houses with low nitrogen dioxide levels and all individuals from houses with high levels.

Computational Exercises

20. Means, Medians, Logs, Ratios. Consider the following tuitions and their natural logs for five colleges:

College	In-State	Out-of-State	Out/In Ratio	Log(In-State)	Log(Out-of-State)
A	\$1,000	\$ 3,000	3	6.9078	8.0064
B	\$4,000	\$ 8,000	2	8.2941	8.9872
C	\$5,000	\$ 30,000	6	8.5172	10.3090
D	\$8,000	\$ 32,000	4	8.9872	10.3735
E	\$40,000	\$ 40,000	1	10.5966	10.5966

(a) Find the average In-State tuition. Find the average log(In-State). Confirm that the log of the average is *not* the same as the average of the logs. (b) Find the median In-State tuition and the median of the logs of In-State tuitions. Verify that the log of the median *is* the same as the median of the logs. (c) Compute the median of the ratios. Compute the differences of logged tuitions—log(Out-of-State) minus log(In-State) and compute the median of these differences. Verify that the median of the differences (of log tuitions) is equal to the natural log of the median of ratios (aside from some minor rounding error).

21. Umpire Life Lengths. When an umpire collapsed and died soon after the beginning of the 1990 U.S. major league baseball season, there was speculation that the stress associated with that job poses a health risk. Researchers subsequently collected historical and current data on umpires to investigate their life expectancies (Cohen et al., “Life Expectancy of Major League Baseball Umpires,” *The Physician and Sportsmedicine*, 28(5) (2000): 83–89). From an original list of 441 umpires, data were found for 227 who had died or had retired and were still living. Of these, dates of birth and death were available for 195. Display 3.10 shows several rows of a generated data set based on the study.

DISPLAY 3.10

First 4 rows (of 227) from the umpire data set (Observed is the known lifetime for those umpires who had died by the time of the study [for whom Censored = 0] and the current age of those who had not yet died [for whom Censored = 1]; Expected is the expected life length—from actuarial life tables—for individuals who were alive at the time the person first became an umpire)

Umpire	Observed life length (yr)	Censored (0 if dead)	Expected life length (yr)
1	63	0	70
2	69	0	71
3	58	0	71
4	61	1	70
...			

- (a) Use a t -test and confidence interval (possibly after transformation) to investigate whether umpires had smaller observed life lengths than expected, using only those with known life lengths (i.e., for whom $Censored = 0$)
- (b) What are the potential consequences of ignoring those 214 of the 441 umpires on the original list for whom data was unavailable?
- (c) What are the potential consequences of ignoring those 32 umpires in the data set who had not yet died at the time of the study? (See, for example, the survival analysis techniques in S. Anderson et al., *Statistical Methods for Comparative Studies*, New York: Wiley, 1980.)

22. Voltage and Insulating Fluid. Researchers examined the time in minutes before an insulating fluid lost its insulating property. The following data are the breakdown times for eight samples of the fluid, which had been randomly allocated to receive one of two voltages of electricity:

Times (min) at 26 kV:	5.79	1579.52	2323.70		
Times (min) at 28 kV:	68.8	108.29	110.29	426.07	1067.60

- (a) Form two new variables by taking the logarithms of the breakdown times: $Y_1 = \log$ breakdown time at 26 kV and $Y_2 = \log$ breakdown time at 28 kV.
- (b) By hand, compute the difference in averages of the log-transformed data: $\bar{Y}_1 - \bar{Y}_2$.
- (c) Take the antilogarithm of the estimate in (b): $\exp(\bar{Y}_1 - \bar{Y}_2)$. What does this estimate? (See the interpretation for the randomized experiment model in Section 3.5.2.)
- (d) By hand, compute a 95% confidence interval for the difference in mean log breakdown times. Take the antilogarithms of the endpoints and express the result in a sentence.

23. Solar Radiation and Skin Cancer. The data in Display 3.11 are yearly skin cancer rates (cases per 100,000 people) in Connecticut, with a code identifying those years that came two years after higher than average sunspot activity and those years that came two years after lower than average sunspot activity. (Data from D. F. Andrews and A. M. Herzberg, *Data*, New York: Springer-Verlag, 1985.) (a) Is there any reason to suspect that using the two independent sample t -test to compare skin cancer rates in the two groups is inappropriate? (b) Draw scatterplots of skin cancer rates versus year, for each group separately. Are any problems indicated by this plot?

24. Sex Discrimination. With a statistical computer program, reanalyze the sex discrimination data in Display 1.3 but use the log transformation of the salaries. (a) Draw box plots. (b) Find a p -value for comparing the distributions of salaries. (c) Find a 95% confidence interval for the ratio of population medians. Write a sentence describing the finding.

DISPLAY 3.11

Partial listing of Connecticut skin cancer rates (per 100,000 people) from 1938 to 1972, with solar code (1 if there was higher than average sunspot activity and 2 if there was lower than average sunspot activity two years earlier)

<u>Year</u>	<u>Rate</u>	<u>Code</u>
1938	0.8	2
1939	1.3	1
1940	1.4	1
1941	1.2	1
...		
1972	4.8	1

DISPLAY 3.12

Proportions of pollen removed and visit durations (in seconds) by 35 bumblebee queens and 12 honeybee workers; partial listing.

<u>Bee</u>	<u>Type</u>	<u>Removed</u>	<u>Duration</u>
1	queen	0.07	2
2	queen	0.10	5
3	queen	0.11	7
4	queen	0.12	11
...			
45	worker	0.78	51
46	worker	0.74	64
47	worker	0.77	78

25. Agent Orange. With a statistical computer program, reanalyze the Agent Orange data of Display 3.3 with and without the two largest dioxin levels in the Vietnam veterans group. Verify the one-sided p -values in bubble 2 of Display 3.7.

26. Agent Orange. With a statistical computer package, reanalyze the Agent Orange data of Display 3.3 after taking a log transformation. Since the data set contains zeros—for which the log is undefined—try the transformation $\log(\text{dioxin} + .5)$. (a) Draw side-by-side box plots of the transformed variable. (b) Find a p -value from the t -test for comparing the two distributions. (c) Compute a 95% confidence interval for the difference in mean log measurements and interpret it on the original scale. (Note: Back-transforming does not provide an exact estimate of the ratio of medians since 0.5 was added to the dioxins, but it does provide an approximate one.)

27. Pollen Removal. As part of a study to investigate reproductive strategies in plants, biologists recorded the time spent at sources of pollen and the proportions of pollen removed by bumblebee queens and honeybee workers pollinating a species of lily. (Data from L. D. Harder and J. D. Thompson, “Evolutionary Options for Maximizing Pollen Dispersal of Animal-pollinated Plants,” *American Naturalist* 133 (1989): 323–44.) Their data appear in Display 3.12.

- (a) (i) Draw side-by-side box plots (or histograms) of the proportion of pollen removed by queens and workers. (ii) When the measurement is the proportion P of some amount, one useful transformation is $\log[P/(1 - P)]$. This is the log of the ratio of the proportion removed to the proportion not removed. Draw side-by-side box plots or histograms on

- this transformed scale. (iii) Test whether the distribution of proportions removed is the same or different for the two groups, using the *t*-test on the transformed data.
- (b) Draw side-by-side box plots of duration of visit on (i) the natural scale, (ii) the logarithmic scale, and (iii) the reciprocal scale. (iv) Which of the three scales seems most appropriate for use of the *t*-tools? (v) Compute a 95% confidence interval to describe the difference in means on the chosen scale. (vi) What are relative advantages of the three scales as far as interpretation goes? (vii) Based on your experience with this problem, comment on the difficulty in assessing equality of population standard deviations from small samples.
- 28. Bumpus's Data.** Obtain *p*-values from the *t*-test to compare humerus lengths for sparrows that survived and those that perished (Exercise 2.21), with and without the smallest length in the perished group (length = 0.659 inch). Do the conclusions depend on this one observation? What action should be taken if they do?
- 29. Cloud Seeding—Multiplicative vs. Additive Effects.** On the computer, create a variable containing the rainfall amounts for only the unseeded days. (a) Create four new variables by adding 100, 200, 300, and 400 to each of the unseeded day rainfall amounts. Display a set of five box plots to illustrate what one might expect if the effect of seeding were additive. (b) Create four additional variables by multiplying each of the unseeded day rainfall amounts by 2, by 3, by 4, and by 5. Display a set of five box plots to illustrate what could be expected if the effect of seeding were multiplicative. (c) Which set of plots more closely resembles the actual data?

Data Problems

- 30. Education and Future Income.** Display 3.13 shows the first five rows of a data set with annual incomes in 2005 of the subset of National Longitudinal Survey of Youth (NLSY79) subjects (described in Exercise 2.22) who had paying jobs in 2005 and who had completed either 12 or 16 years of education by the time of their interview in 2006. All the subjects in this sample were between 41 and 49 years of age in 2006. Analyze the data to describe the amount (or percent) by which the population distribution of incomes for those with 16 years of education exceeds the distribution for those with 12 years of education. (*Note:* The NLSY79 data set codes all incomes above \$150,000 as \$279,816. To make an exercise version that better matches the actual income distribution, those values have been replaced in the data set by computer-simulated values from a realistic distribution of incomes greater than \$150,000.)

DISPLAY 3.13

Annual incomes in 2005 (in U.S. dollars) of 1,020 Americans who had 12 years of education and 406 who had 16 years of education by the time of their interview in 2006; "Subject" is a subject identification number; first 5 of 1,426 rows

Subject	Educ	Income2005
2	12	5,500
6	16	65,000
7	12	19,000
13	16	8,000
21	16	253,043

- 31. Education and Future Income II.** The data file ex0331 contains a subset of the NLSY79 data set (see Exercise 30) with annual incomes of subjects with either 16 or more than 16 years of

education. Analyze the data to describe the amount (or percent) by which the population distribution of incomes for those with more than 16 years of education exceeds the distribution for those with 16 years of education.

32. College Tuition. Display 3.14 shows the first five rows of a data set with 2011–2012 in-state and out-of-state tuitions for random samples of 25 private and 25 public, four-year colleges and universities in the United States. Analyze the data to describe (a) the extent to which out-of-state tuition is more expensive than in-state tuition in the population of public schools, (b) the extent to which private school in-state tuition is more expensive than public school in-state tuition, and (c) the extent to which private school out-of-state tuition is more expensive than public school out-of-state tuition. (Data sampled from College Board: <http://www.collegeboard.com/student/> (11 July 2011).)

DISPLAY 3.14

In-state and out-of-state tuitions for 25 public and 25 private colleges and universities in the United States; first 5 of 50 rows

College	Type	InState	OutofState
Albany State University	public	\$5,434	\$17,048
Appalachian State University	public	\$5,175	\$16,487
Argosy University: Nashville	private	\$19,596	\$19,596
Brescia University	private	\$18,140	\$18,140
Central Connecticut State University	public	\$8,055	\$18,679

33. Brain Size and Litter Size. Display 3.15 shows relative brain weights (brain weight divided by body weight) for 51 species of mammal whose average litter size is less than 2 and for 45 species of mammal whose average litter size is greater than or equal to 2. (These are part of a larger data set considered in Section 9.1.2.) What evidence is there that brain sizes tend to be different for the two groups? How big of a difference is indicated? Include the appropriate statistical measures of uncertainty in carefully worded sentences to answer these questions.

DISPLAY 3.15

Relative brain sizes, $1,000 \times (\text{Brain weight}/\text{Body weight})$, for 96 species of mammals

$1,000 \times (\text{Brain weight}/\text{Body weight})$ for 51 species with average litter size < 2

0.42	0.86	0.88	1.11	1.34	1.38	1.42	1.47	1.63	1.73	2.17	2.42
2.48	2.74	2.74	2.79	2.90	3.12	3.18	3.27	3.30	3.61	3.63	4.13
4.40	5.00	5.20	5.59	7.04	7.15	7.25	7.75	8.00	8.84	9.30	9.68
10.32	10.41	10.48	11.29	12.30	12.53	12.69	14.14	14.15	14.27	14.56	15.84
18.55	19.73	20.00									

$1,000 \times (\text{Brain weight}/\text{Body weight})$ for 45 species with average litter size ≥ 2

0.94	1.26	1.44	1.49	1.63	1.80	2.00	2.00	2.56	2.58	3.24	3.39
3.53	3.77	4.36	4.41	4.60	4.67	5.39	6.25	7.02	7.89	7.97	8.00
8.28	8.83	8.91	8.96	9.92	11.36	12.15	14.40	16.00	18.61	18.75	19.05
21.00	21.41	23.27	24.71	25.00	28.75	30.23	35.45	36.35			

Answers to Conceptual Exercises

1. The target clouds on a day that was deemed suitable for seeding.
2. Uncontrollable confounding factors probably explain the variability in rainfall from clouds treated the same way. Randomization is needed to ensure that the confounding factors do not tend to be unevenly distributed in the two groups.
3. Blinding prevents the intentional or unintentional biases of the human investigators from having a chance to make a difference in the results.
4. There may be serial correlation. A plot of rainfall versus date could be used to check.
5. Larger values are to be expected by chance if the populations are the same, since the sample of Vietnam veterans is so much larger than the sample of non-Vietnam veterans.
6. (a) He would not be representative of the target population and should be removed from the data set for analysis. (b) Same thing.
7. No, not from the statistics alone since this is an observational study. It could be said, however, that the data are consistent with that theory.
8. No. The dependence is the result of matching and is desirable. The two-sample t -tools are not appropriate (but the paired t -tools are).
9. False.
10. False. An *average* from a sample will have a sampling distribution that will tend toward normal with large sample sizes, but the sample histogram should mirror the population distribution. As the sample size gets larger, the sample histogram should become a better approximation to the population histogram.
11. There is a cluster effect: the particular day of the week. She should use a paired- t analysis, as will be discussed in Chapter 4.
12. The t -test is robust in validity to departures from normality, especially as the sample size gets large. The robustness with respect to equal standard deviations does not depend much on what the sample sizes are, so long as they are reasonably equal. Sample size does not affect robustness with respect to independence.
13. You cannot take logarithms of negative numbers.
14. It will be more consequential in smaller samples; its effect gets washed out in large ones.
15. Replace the population SDs in the formula (Section 2.2.2) by individual *sample* SDs.
16. No. The paired analysis must be used, even though the inferences may not appear to be as precise. The unpaired analysis is inappropriate.
17. (a) Dependence of measurements on individuals in the same household (cluster effect). (b) Maybe. Getting a single measure for each household may be an easy way out of the dependence problem, but care should be used as these groups also tend to differ in the average number of persons per household.
18. (a) No. (b) Yes.
19. (a) Serial correlation both within and between groups. (Confounding variable is the time at which observations were made.) (b) Cluster effect between groups. (c) Cluster effect (members of the same household should be similar) within groups.

Alternatives to the *t*-Tools

The *t*-tools have an extremely broad range of application, extending well beyond the strict confines of the ideal model because of robustness. They extend even further when the possibilities of transforming the data and dealing with outliers are taken into account.

Nevertheless, situations arise where the *t*-tools cannot be applied, because the model assumptions of the *t*-test are grossly violated. For these situations, a host of other methods, based on different models, may be used. Some are presented in this chapter. Most notable are two distribution-free methods, based on models that do not specify any particular population distributions. The rank-sum test for two independent samples and the signed-rank test for a sample of pairs are useful alternatives, particularly when outliers may be present or when the sample sizes are too small to permit the assessment of distributional assumptions.

4.1 CASE STUDIES

4.1.1 Space Shuttle O-Ring Failures—An Observational Study

On January 27, 1986, the night before the space shuttle *Challenger* exploded, engineers at the company that built the shuttle warned National Aeronautics and Space Administration (NASA) scientists that the shuttle should not be launched because of predicted cold weather. Fuel seal problems, which had been encountered in earlier flights, were suspected of being associated with low temperatures. It was argued, however, that the evidence was inconclusive. The decision was made to launch, even though the temperature at launch time was 29°F.

The data in Display 4.1 are the numbers of O-ring incidents on previous shuttle flights, categorized into those launched at temperatures below 65°F and those launched at temperatures above 65°F. (Data from a graph in Richard P. Feynman, *What Do You Care What Other People Think?* (New York: W. W. Norton, 1988).) Is there a higher risk of O-ring incidents at lower launch temperatures?

DISPLAY 4.1 Numbers of O-ring incidents on 24 space shuttle flights prior to the *Challenger* disaster

Launch temperature	Number of O-ring incidents														
Below 65°F	1	1	1	3											
Above 65°F	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1

Statistical Conclusion

These date provide strong evidence that the number of O-ring incidents was associated with launch temperature in these 24 launches. It is highly unlikely that the observed difference of the groups is due to chance (one-sided p -value = 0.0099 from a permutation test on the t -statistic).

Scope of Inference

These observational data cannot be used to establish causality, nor is there any broader population of which they are a sample. But the association between temperature and O-ring failure in these particular 24 launches is consistent with the theory that lower temperatures impair the functioning of the O-rings. (At one point in public hearings into the causes of the disaster, Feynman asked for a glass of ice water, placed a small O-ring in it for a time, removed it, and then proceeded to demonstrate that the rubber failed to spring back to its original form.) (Note: Other techniques for dealing with count data are given in Chapter 22.)

4.1.2 Cognitive Load Theory in Teaching—A Randomized Experiment

Consider the following problem in coordinate geometry.

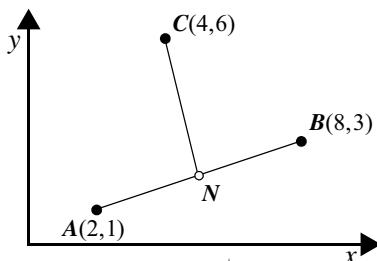
Point A has coordinates $(2, 1)$, point B has $(8, 3)$, and point C has $(4, 6)$.

What is the slope of the line that connects C to the midpoint between A and B ?

Presenting the solution as a worked problem, a conventional textbook shows a picture of the layout, gives a discussion in the text, and then provides the lines of algebraic manipulation leading to the right answer. (See Display 4.2.) Recent theoretical developments in cognitive science suggest that splitting the presentation into the three distinct units of diagram, text, and algebra imposes a heavy, extraneous cognitive load on the student. The requirement that the student organize and process the separate elements constitutes a cognitive load. The load is extraneous because it is not essential to learning how to solve such problems—indeed, it impedes the learning process by placing heavy demands on cognitive resources that should be used to understand the essentials.

DISPLAY 4.2

Cognitive load experiment: conventional method of instruction (for finding the slope of the line that connects C to the midpoint between A and B)



Solution: The coordinates of N are:

$$N = \left(\frac{2+8}{2}, \frac{1+3}{2} \right) \\ = (5, 2)$$

The slope of CN is:

$$m = \frac{2-6}{5-4} \\ = \frac{-4}{1} = -4$$

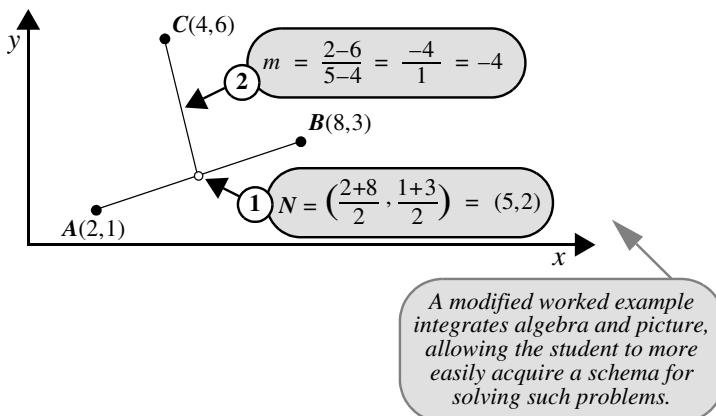
In a conventional worked example, algebra and diagram are separated, giving students an extraneous cognitive load of having to assimilate the two.

In a test of this theory, researchers compared the effectiveness of conventional textbook worked examples to modified worked examples, which present the algebraic manipulations and explanation as part of the graphical display (see Display 4.3). (Data from J. Sweller, P. Chandler, P. Tierney, and M. Cooper, “Cognitive Load as a Factor in the Structuring of Technical Material,” *Journal of Experimental Psychology General* 119(2) (1990): 176–92.)

Researchers selected 28 ninth-year students in Sydney, Australia, who had no previous exposure to coordinate geometry but did have adequate mathematics to deal with the problems given. The students were randomly assigned to one of

DISPLAY 4.3

Cognitive load experiment: modified method of instruction (for finding the slope of the line that connects C to the midpoint between A and B)



two self-study instructional groups, using conventional and modified instructional materials. The materials covered exactly the same problems, presented differently. Students were allowed as much time as they wished to study the material, but they were not allowed to ask questions. Following the instructional phase, all students were tested with a common examination over three problems of different difficulty. The data in Display 4.4, based on this study, are the number of seconds required to arrive at a solution to the moderately difficult problem.

Both distributions in Display 4.4 are highly skewed. In addition, there were five students in the conventional (control) group who did not come to any solution in the five minutes allotted. Their solution times are considered *censored*—all that is known about them is that they exceed 300 seconds. It appears that the solution times for the “modified instructional materials” group are generally shorter than for the conventional materials group. Is there sufficient evidence to draw this conclusion?

Statistical Conclusions

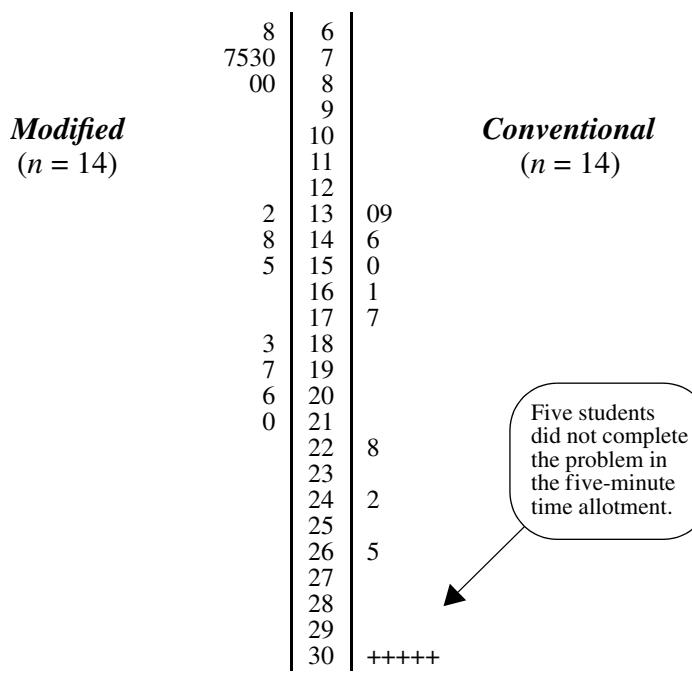
These data provide convincing evidence that a student could solve the problem more quickly if taught with the modified method than if taught with the conventional method (one-sided p -value = 0.0013, from the rank-sum test). The modified instructional materials shortened solution times by an estimated 129 seconds (95% confidence interval for an additive treatment effect: 58 to 159 seconds).

4.2 THE RANK-SUM TEST

The *rank-sum test* is a resistant alternative to the two-sample *t*-test. It performs nearly as well as the *t*-test when the two populations are normal and considerably

DISPLAY 4.4

Number of seconds to solution of a problem in coordinate geometry, for students instructed with conventional and modified materials



Five students did not complete the problem in the five-minute time allotment.

Legend: | 26 | 5 represents 265 seconds.

better when there are extreme outliers. Its drawbacks are that associated confidence intervals are not computed by most of the statistical computer packages and that it does not easily extend to more complicated situations.

4.2.1 The Rank Transformation

Straightforward transformations of the data were used in Chapter 3 to obtain measurements on a scale where the normality and equal spread assumptions are approximately met. The rank-sum relies on a special kind of transformation that replaces each observation by its rank in the combined sample. It should be noted that this really is a different kind of transformation for two reasons: (1) A single transformed value depends on all the data. So the transformed rank of, say, $Y = 63.925$, may be one value in one problem but a very different value in another problem. (2) There is no reverse transformation, such as the antilog, available for ranks.

The purpose behind using ranks is not to transform the data to approximate normality but to transform the data to a scale that eliminates the importance of the

population distributions altogether. Notice that when the data values are replaced by their square roots or by their logarithms, the sample distributions change shape dramatically. The ranks from the transformed data, however, are identical to the ranks based on the untransformed data, so the shape of the distribution of measurements has no effect on the ranks. In addition, whether the largest observation in a data set is 300 seconds or 3,000 seconds does not affect its rank. In this respect, any statistic based on ranks is resistant to outliers.

A feature of the rank-sum test that makes it an attractive choice for the cognitive load experiment is its ability to deal with *censored observations*, observations that are only known to be greater than (or possibly less than) some number. All that is known about the five students who did not complete the problem is that their solution times are greater than 300 seconds. For the rank-sum test it is enough to know that, in terms of ranks, they were tied for last.

4.2.2 The Rank-Sum Statistic

Calculation of the rank-sum test statistic for the cognitive load experiment is summarized in Display 4.5. The first four steps transform the data to their ranks in the combined sample:

1. List all observations from both samples in increasing order.
2. Identify which sample each observation came from.
3. Create a new column labeled “order,” as a straight sequence of numbers from 1 to $(n_1 + n_2)$.
4. Search for ties—that is, duplicated values—in the combined data set. The ranks for tied observations are taken to be the average of the orders for those cases.

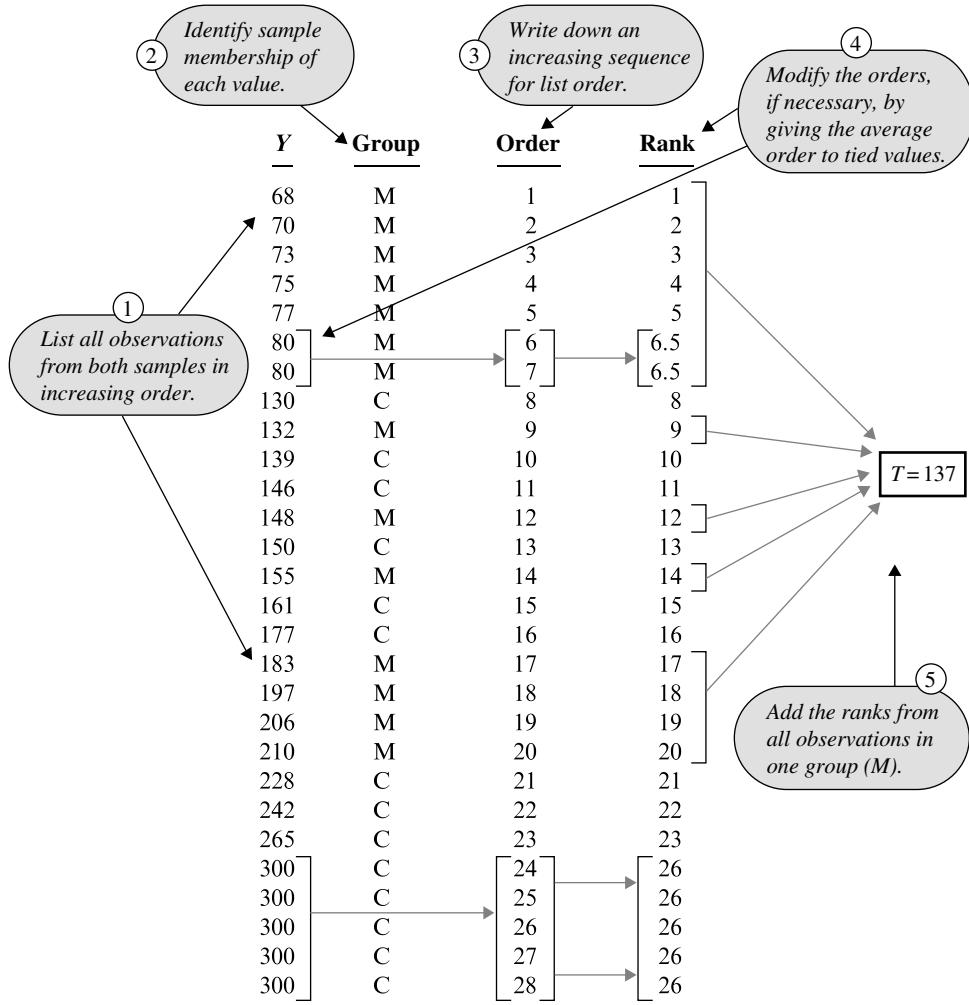
Two students tied at 80 seconds, for example. They finished sixth and seventh fastest. One cannot say which of the two deserves which order, so both are assigned the rank of $(6 + 7)/2 = 6.5$. The five students who took the full five minutes were the last five finishers, with orders 24, 25, 26, 27, and 28. So each is assigned rank $(24 + 25 + 26 + 27 + 28)/5 = 26$. Any observation that has a unique value gets its order as its rank.

The test statistic, T , is the sum of all the ranks in one group, called “group 1.” Group 1 is conventionally the group with the smaller sample size (because that minimizes computation). The choice, however, is arbitrary.

The impressive feature of the cognitive load experiment is that the early finishers were mostly the students who studied the modified instructional material. This is reflected in the test statistic by a low rank-sum ($T = 137$) for that group.

4.2.3 Finding a *p*-Value by Normal Approximation

The rank-sum procedure is used to test the null hypothesis of no treatment effect from a two-treatment randomized experiment and also to test the null hypothesis of identical population distributions from two independent samples. If the null hypothesis is true in either case, then the sample of n_1 ranks in group 1 is a random sample from the $n_1 + n_2$ available ranks. Both the randomization distribution and the sampling distribution of the rank-sum statistic, T , are represented by a

DISPLAY 4.5Rank-sum test statistic T for the cognitive load experiment

histogram of the rank-sum statistics calculated for each possible regrouping of the $n_1 + n_2$ observations into samples of size n_1 and n_2 .

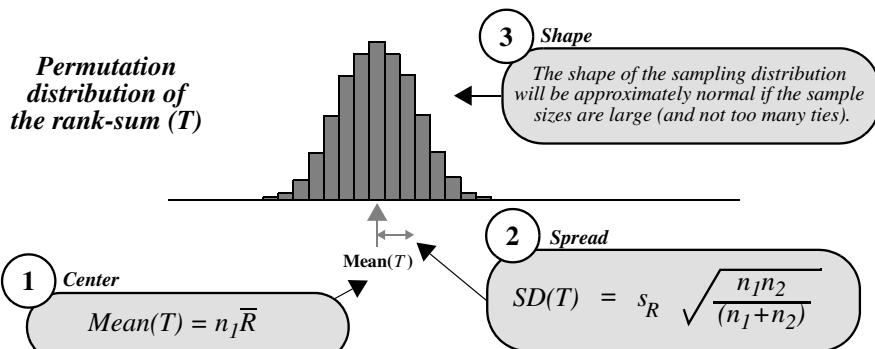
Because conversion to ranks avoids absurd distributional anomalies, the randomization distribution of T can be approximated accurately by a normal distribution in most situations. The exceptions are when at least one sample is small—say under 5—or when large numbers of ties occur. The mean and variance of the permutation distribution, from statistical theory, are shown in Display 4.6.

These facts may be used to evaluate whether the observed rank-sum statistic is unusually small or large. If there is no difference, then the Z -statistic

$$Z\text{-statistic} = [T - \text{Mean}(T)]/\text{SD}(T)$$

DISPLAY 4.6

Facts about the randomization (or sampling) distribution of the rank-sum statistic—the sum of ranks in group 1—when there is no group difference



where \bar{R} and s_R are the average and the sample standard deviation, respectively, for the combined set of $(n_1 + n_2)$ ranks.

should be similar to a typical value from a standard normal distribution. A *p*-value is the proportion of values from a standard normal distribution that are more extreme than the observed *Z*-statistic.

The calculations for the cognitive load data are shown in Display 4.7. The observed rank-sum statistic of 137 is smaller than the value of 203 that would be expected if there were no treatment effect. This disparity may be due to the luck involved in the random assignment, with better students placed in the modified instruction group, but the *p*-value indicates that only about one in a thousand randomizations would produce a disparity as great as or greater than the observed one, if there were no treatment effect. Such a small *p*-value provides fairly convincing evidence that a treatment effect explains the difference.

Continuity Correction

The statistic T can only take on integer values. Its distribution can be displayed as a histogram with a bar centered on each integer having a height equal to the probability that T equals that integer and a base extending from the integer minus one-half to the integer plus one-half. Thus the bar areas are the probabilities associated with each integer. Probabilities can also be approximated by the normal distribution, which is centered at $\text{Mean}(T)$ with a standard deviation of $\text{SD}(T)$.

Using the normal distribution to approximate the probability that T is less than or equal to the integer k , one must realize that the desired area under the normal curve is the area less than or equal to $k + 0.5$ in the probability histogram. To get the best approximation, therefore, one must determine the normal distribution area less than or equal to $k + 0.5$.

DISPLAY 4.7

Finding the p -value with the normal approximation to the permutation distribution of the rank-sum statistic, using a continuity correction. Calculations for the cognitive load data are continued from Display 4.5

- 1** Calculate the average and sample standard deviation of the ranks from the combined sample (column 4 of Display 4.5).

$$\bar{R} = 14.5 \quad s_R = 8.2023$$

- 2** Compute the theoretical “null hypothesis” mean and standard deviation of T using the formulae in Display 4.6.

$$\text{Mean}(T) = 14 \times 14.5 = 203 \quad \text{SD}(T) = 8.2023 \sqrt{\frac{14 \times 14}{(14 + 14)}} = 21.7013$$

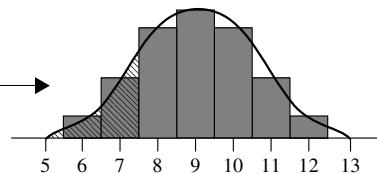
- 3** Calculate the Z-statistic using a continuity correction.

$$Z = \frac{(137.5 - 203)}{21.7013} = -3.0183$$

- 4** Find the p -value from a standard normal table. **One-sided p -value = 0.0013**

Continuity Correction

The probability (shaded area) in the histogram for values ≤ 7 is approximated by normal curve area (hatched) for values ≤ 7.5 .



Similarly, the normal approximation to the probability that T is greater than or equal to an integer k is the area under the normal curve to the right of $k - 0.5$ (so that the entire bar above k is included).

This adjustment to the normal approximation is called a *continuity correction* because it corrects the less accurate calculation that simply uses k . The calculations in Display 4.7 include this feature.

Exact p -Values for the Rank-Sum Test

The normal distribution is an easy and adequate approximation to the randomization distribution for most problems. Troublesome situations arise where sample sizes are small or there are large numbers of ties.

It is possible to compute the randomization distribution exactly using the technique described in Section 1.3.2. Published tables with exact percentiles exist (see, for example, Pearson and Hartley 1972, p. 227), but they are only appropriate when

there are no ties. Another drawback is that only a few percentiles are published for each combination of sample sizes, so that *p*-values can only be bracketed. Many statistical computer programs can compute the exact *p*-value from the permutation distribution (if the sample sizes aren't too large) or approximate it through repeated sampling from the permutation distribution.

4.2.4 A Confidence Interval Based on the Rank-Sum Test

A 95% confidence interval for any parameter can be constructed by including all hypothesized values that lead to two-sided *p*-values greater than or equal to 0.05. This relation between tests and confidence intervals can be exploited to obtain a confidence interval for an additive treatment effect δ in the cognitive load experiment.

If Y is the time to solution for a student using the conventional study materials, the additive treatment effect model says that $Y - \delta$ is the time to solution for the same student using the modified study materials. The rank-sum test in Display 4.7 indicates that 0 is not very likely as a value for δ . Could δ be, say, 50 seconds? If so, the completion times in the modified group *plus 50 seconds* should be a set of times that are similar to those for the conventional study materials group. This suggests a procedure: (1) add a hypothesized δ to all modified group times, (2) use the rank-sum test to decide whether the two resulting group differences can be explained by chance (in this case whether the two-sided *p*-value exceeds 0.05), and (3) determine—through trial and error—upper and lower limits on δ values satisfying the criterion.

Display 4.8 illustrates the process with a series of proposed values for δ . By trial and error it was determined that all hypothesized values of δ between 58 and 159 seconds lead to two-sided *p*-values greater than 0.05. The 95% confidence interval for the reduction in test time due to the modified instructional method is 58 to 159 seconds. A point estimate for δ is the interval's midpoint, 108 seconds.

DISPLAY 4.8

Using a rank-sum test to construct a confidence interval for an additive treatment effect (cognitive load study)

Hypothesized effect (seconds)	Two-sided <i>p</i> -value	Confidence interval inclusion?
50	0.0286	no
60	0.0800	yes
55	0.0403	no
58	0.0502	yes
150	0.1227	yes
160	0.0476	no
155	0.0589	yes
158	0.0530	yes
159	0.0502	yes

Try several hypothesized values for δ to identify those that have two-sided *p*-values ≥ 0.05 .

A 95% confidence interval is –159 seconds to –58 seconds.

Notes About the Rank-Sum Procedure

1. Other names for the rank-sum test are the Wilcoxon test and the Mann–Whitney test. The different names refer to originators of different forms of the test statistic. To confuse the issue, there is also a Wilcoxon signed-rank test—which is something entirely different.
2. The rank-sum test is a *nonparametric* or *distribution-free* statistical tool, meaning there are no specific distributional assumptions required.
3. Although the *t*-test is more efficient when the populations are normal (it makes better use of the available data), the rank-sum test is not that much worse in the normal model, and is substantially better for many other situations, particularly for long-tailed distributions.
4. The theoretical mean of T in Display 4.6 can also be written as $\text{Mean}(T) = n_1(n_1 + n_2 + 1)/2$. If there are no ties, the theoretical standard deviation in the permutation distribution of T is

$$\text{SD}(T) = \sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}$$

The version in bubble 2 of Display 4.6 is correct whether there are ties or not.

5. In its application to the cognitive load problem, the confidence interval method described in Display 4.8 should be modified to reflect the truncation of solution times to 5 minutes. Thus, when the modified group's times are shifted by a certain hypothesized amount, any values exceeding 300 seconds should be replaced by 300 before performing the test. Such a modification makes no change in the lower limit of the confidence interval; but the upper limit changes from 159 to 189 seconds.

4.3 OTHER ALTERNATIVES FOR TWO INDEPENDENT SAMPLES

4.3.1 Permutation Tests

A *permutation test* is any test that finds a p -value as the proportion of regroupings—of the observed $n_1 + n_2$ numbers into two groups of size $n_1 + n_2$ —that lead to test statistics as extreme as the observed one. The test statistic may be the difference in group averages, a *t*-statistic, the sum of ranks in group 1, or any other choice to represent group difference. Permutation tests were previously discussed in the context of interpretation. When used to analyze randomized experiments, for example, permutation tests are called *randomization tests* and provide statistical inferences tied to the chance mechanism in random assignment. For observational studies, they provide no inference to a broader context (except for the special case of the rank-sum test), but may nevertheless be useful for summarizing differences in the data at hand. In the cases so far discussed, the actual calculation of p -values and confidence intervals was based on an approximation to the permutation distribution. Sometimes there is no adequate approximation.

In the O-ring study, for example, the exact rendering of the permutation test is the only method for calculating its p -value. The distribution of the numbers is so

nonnormal that a *t*-distribution approximation is severely inadequate. No transformation helps. The rank-sum test is inadequate because of the large number of ties: 17 of the 24 values are tied at zero. Even though the computational effort for direct calculation of the *p*-value is considerable, permutation calculations are important because they are always available and require no distributional assumptions or special conditions. The *p*-value is calculated using the following procedure:

1. Decide on a test statistic, and compute its value from the two samples.
2. List all regroupings of the $n_1 + n_2$ numbers into groups of size n_1 and n_2 , and recompute the test statistic for each.
3. Count the number of regroupings that produce test statistics at least as extreme as the observed test statistic from step 1.
4. The *p*-value is the number found in step 3 divided by the total number of regroupings.

In problems such as the O-ring study, the procedure can be accomplished by counting with *combinatorics*, that is, by using the *combination numbers*

$$C_{n,k} = \frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 1}.$$

The number $C_{n,k}$ —read as “ n choose k ”—is the number of different ways to choose k items from a list of n items. The total number of regroupings for a two-sample problem is $C_{n_1+n_2,n_1}$. Combination numbers can also be used to count the number of regroupings that lead to test statistics as extreme or more extreme than the observed one, as now shown for the O-ring data.

Step 1. The *t*-statistic was selected as the test statistic (although the difference in averages would be equally good). Its observed value is 3.888.

Step 2. It is only necessary to determine which group 1 outcomes produce *t*-statistics that are as large as or larger than the 3.888 that came from the observed outcome, (1, 1, 1, 3). After some calculation, it is found that the extreme regroupings are those whose group 1 outcomes are (1, 1, 2, 3), (1, 1, 1, 3), or (0, 1, 2, 3), with *t*-statistics 5.952, 3.888, and 3.888, respectively.

Step 3. The total number of regroupings is $C_{24,4} = 10,626$. For a one-sided *p*-value it is necessary to count the number of regroupings that produce the outcomes in Step 2. Consider the outcome (1, 1, 2, 3). To get such an outcome, the single 2 and the single 3 must be selected. However, there are five 1's in the full sample, and the indicated outcome would occur with any of the $C_{5,2} = 10$ combinations of 1's. This is illustrated in Display 4.9.

Similarly, the number of regroupings with outcome (1, 1, 1, 3) is the number of ways to select three 1's from the five available in the data set, to go with the obligatory 3. This is $C_{5,3} = 10$. Finally, a regrouping with outcome (0, 1, 2, 3) can be formed by taking any one of the 17 0's in combination with any one of the five 1's, along with 2 and 3. The number of ways to do this is $C_{17,1} \times C_{5,1} = 17 \times 5 = 85$.

DISPLAY 4.9

Counting method for regroupings that give extreme group 1 outcomes

DISPLAY 4.10

A summary of the t -statistics calculated from all 10,626 rearrangements of the O-ring data into a Low group of size 4 and a High group of size 20

Number of rearrangements with identical t-statistics	t-statistic	<i>Total number of rearrangements into two groups of size 4 and 20:</i>
2,380	-1.188	10,626
3,400	-0.463	
2,040	0.231	
1,530	0.939	
855	1.716	
316	2.643	
95	3.888	<i>Number of rearrangements with t-statistics greater than or equal to 3.888:</i>
10	5.952	105
		<i>one-sided p-value from a permutation test of the t-statistic:</i>
		105/10,626 = 0.00988

Summing up, the number of regroupings that lead to t -statistics as large as or larger than the observed one is $10 + 10 + 85 = 105$.

Step 4. The one-sided p -value is $105/10,626 = 0.00988$. This result and the full permutation distribution of the t -statistic appear in Display 4.10.

Notes: The combinatorics method is especially useful if there are only a few regroupings that need to be counted, but it might be unmanageable if there are many. A computer can systematically enumerate the regroupings, calculate the test statistic for each, and compute the proportions that have a more extreme test statistic than the observed one. (An alternative is to approximate the p -value by the proportion of a *random sample* of all possible regroupings that have a test statistic more extreme than the observed one, as discussed for the discrimination study in Section 1.1.2.)

4.3.2 The Welch t -Test for Comparing Two Normal Populations with Unequal Spreads

Welch's t-test employs the individual sample standard deviations as separate estimates of their respective population standard deviations, rather than pooling to

obtain a single estimate of a population standard deviation. The result is a different formula for the standard error of the difference in averages:

$$\text{SE}_W(\bar{Y}_2 - \bar{Y}_1) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

This becomes the denominator in the *t*-statistic for comparing the means of populations with different spreads. Even when the populations are normal, however, the exact sampling distribution of the Welch *t*-ratio is unknown. It can be approximated by a *t*-distribution with d.f._W degrees of freedom, known as Satterthwaite's approximation:

$$\text{d.f.}_W = \frac{\frac{[\text{SE}_W(\bar{Y}_2 - \bar{Y}_1)]^4}{[\text{SE}(\bar{Y}_2)]^4} + \frac{[\text{SE}(\bar{Y}_1)]^4}{(n_2 - 1)}}{(n_2 - 1) + \frac{[\text{SE}(\bar{Y}_1)]^4}{(n_1 - 1)}}$$

where

$$\text{SE}(\bar{Y}_1) = \frac{s_1}{\sqrt{n_1}} \quad \text{and} \quad \text{SE}(\bar{Y}_2) = \frac{s_2}{\sqrt{n_2}}.$$

The *t*-test and confidence interval are computed exactly as with the two-sample *t*-test, except with the modified standard error and the approximate degrees of freedom (rounded down to an integer value).

A Note About the Importance of the Equal-Spread Model

If the two populations have the same shape and the same spread, then the difference in means is an entirely adequate summary of their difference. Any question of interest that requires a comparison of the two distributions can be re-expressed in terms of the single parameter $\mu_2 - \mu_1$.

If, on the other hand, the populations have different means *and different standard deviations*, then $\mu_2 - \mu_1$ may be an inadequate summary. If lifetimes of brand A lightbulbs have a larger mean and a larger standard deviation than lifetimes of brand B lightbulbs, as shown in Display 4.11, then there is a higher proportion of long-life bulbs from brand A, but there may also be a higher proportion of short-life bulbs from brand A. A comparison of brands is not entirely resolved by a comparison of means.

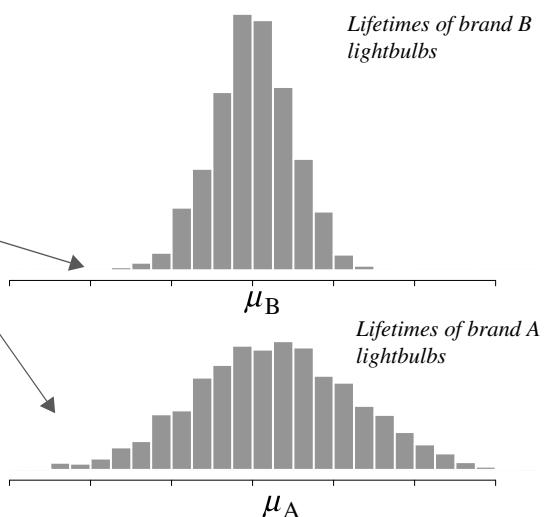
Although some statistical analysts find the unequal variance model more appealing for the two-sample problem, most of the standard methods for more complicated structures make use of a pooled estimate of variance. In that sense, the two-sample *t*-tools extend to more complicated situations more easily than does Welch's *t*-test.

DISPLAY 4.11

The conceptual difficulty with comparing population means when population spreads are not the same

On average, brand A bulbs last longer; but there is also a greater chance of early burnout with brand A.

The question of which brand is better may be more complex than simply “Which mean is larger?”



4.4 ALTERNATIVES FOR PAIRED DATA

Two resistant and distribution-free alternatives to the paired *t*-test are described here. The sign test is a tool for a quick analysis of paired data. If the sign test shows a substantial effect, that may be enough to resolve the answer to the question of interest. But it is not very efficient. If its results are inconclusive, the signed-rank test is a more powerful alternative.

4.4.1 The Sign Test

The *sign test* is a quick, distribution-free test of the hypothesis that the mean difference of a population of pairs is zero (for observational studies) or of the hypothesis that there is no treatment effect in a randomized paired experiment. It counts the number K of pairs where one group's measurement exceeds the other's. In the schizophrenia study (Section 2.1.2), the left hippocampus volume of the unaffected twin was larger than the left hippocampus volume of the affected twin for $K = 14$ out of the $n = 15$ twin pairs. If there were no systematic difference between affected and unaffected individuals, one would expect K to be near $n/2$, as both groups share equal chances for having the larger measurement. The sign test provides evidence against the null hypothesis whenever K is far from $n/2$.

If the null hypothesis is true, the distribution of K is *binomial*, indexed by $n = 15$ trials and probability 0.5 of a positive difference in each trial. The chance of obtaining exactly k positive differences is given by $C_{n,k} \times (1/2)^n$ (with $C_{n,k}$ as defined in Section 4.3.1). The *p*-value is the sum of these chances for all values of k that are as far or farther from $n/2$ than is the observed value K .

An approximation to the *p*-value comes from normal approximation to the binomial distribution. If the null hypothesis is true, then the *Z*-statistic

$$Z\text{-statistic} = \frac{K - (n/2)}{\sqrt{n/4}}$$

has approximately a standard normal distribution. A *p*-value is found as the proportion of a standard normal distribution that is farther from zero than the observed *Z*-statistic. The approximation can be improved with a continuity correction, by adding $\frac{1}{2}$ to (or subtracting it from) the numerator of the *Z*-statistic, to make it closer to zero. (*Note:* n is the number of observations with nonzero differences; we discard ties in using the sign test.)

Example—Schizophrenia

Since 14 of the 15 differences are positive, the exact one-sided *p*-value is the probability that k is 14 plus the probability that k is 15. This is $C_{15,14} \times (1/2)^{15} + C_{15,15} \times (1/2)^{15} = 15 \times (1/32,768) + 1 \times (1/32,768) = 0.00049$. To illustrate the normal approximation, the *Z*-statistic with continuity correction is

$$Z\text{-statistic} = \frac{[14 - (15/2)] - (1/2)}{\sqrt{15/4}} = 3.098,$$

which produces an approximate one-sided *p*-value of 0.00097.

4.4.2 The Wilcoxon Signed-Rank Test

By retaining only the signs of the difference for each pair, the sign test completely obliterates the effects of outliers, but at the expense of losing potentially useful information about the relative magnitudes of the differences. The *Wilcoxon signed-rank test* uses the ranks of the magnitudes of the differences in addition to their signs. Since ranks are used, the procedure is resistant to outliers.

Computation of the *signed-rank statistic* proceeds as follows.

1. Compute the difference in each of the n pairs.
2. Drop zeros from the list.
3. Order the *absolute differences* from smallest to largest and assign them their ranks $1, \dots, n$ (or average rank for ties).
4. The signed-rank statistic, S , is the sum of the ranks from the pairs for which the difference is positive.

The computation of S is illustrated in the schizophrenia data in Display 4.12, for the hypothesis that the mean difference is zero.

DISPLAY 4.12 Signed-rank test statistic computations (schizophrenia study)

Pair	Unaffected	Affected	Difference	Ordered magnitude	Order	Rank	+ Ranks	- Ranks
1	1.94	1.27	0.67	0.02 (+)	1	1	1	
2	1.44	1.63	-0.19	0.03 (+)	2	2	2	
3	1.56	1.47	0.09	0.04 (+)	3	3	3	
4	1.58	1.39	0.19	0.07 (+)	4	4	4	
5	2.06	1.93	0.13	0.09 (+)	5	5	5	
6	1.66	1.26	0.40	0.10 (+)	6	6	6	
7	1.75	1.71	0.04	0.11 (+)	7	7	7	
8	1.77	1.67	0.10	0.13 (+)	8	8	8	
9	1.78	1.28	0.50	0.19 (+)	9	9.5	9.5	
10	1.92	1.85	0.07	0.19 (-)	10	9.5		9.5
11	1.25	1.02	0.23	0.23 (+)	11	11	11	
12	1.93	1.34	0.59	0.40 (+)	12	12	12	
13	2.04	2.02	0.02	0.50 (+)	13	13	13	
14	1.62	1.59	0.03	0.59 (+)	14	14	14	
15	2.08	1.97	0.11	0.67 (+)	15	15	15	
							<u>= 110.5</u>	

① Order the absolute differences and assign ranks to them.

② Signed-rank statistic = Sum of ranks for positive differences.

Exact *p*-Value

An exact *p*-value for the signed-rank test is the proportion of all assignments of outcomes within each pair that lead to a test statistic as extreme as or more extreme than the one observed. Assignment refers to switching the group status but keeping the same observations within each pair. Within a single pair there are two possible assignments, so with n pairs there are a total of 2^n possible assignments. The *p*-value is therefore the number of possible assignments that provide a sum of positive ranks as extreme as or more extreme than the observed one, divided by 2^n .

For the schizophrenia example, there are 15 pairs, so the labels “affected” and “unaffected” can be assigned to observations within pairs in $2^{15} = 32,768$ possible ways. The assignments that produce an *S* greater than or equal to 110.5 are those where the sum of the negative ranks is less than or equal to 9.5.

By systematically examining the possible assignments with no negative ranks, one negative rank, two negative ranks, and so on, the following assignments (indicated by the ranks associated with a negative sign) are found to have a sum of negative ranks less than or equal to 9.5: (none), (1), (2), (3), (4), (5), (6), (7), (8), (9.5), (9.5), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (1, 7), (1, 8), (2, 3), (2, 4), (2, 5), (2, 6), (2, 7), (3, 4), (3, 5), (3, 6), (4, 5), (1, 2, 3), (1, 2, 4), (1, 2, 5), (1, 2, 6), (1, 3, 4), (1, 3, 5), and (2, 3, 4). There are 34 assignments with the sum of negative ranks less than or

equal to 9.5, and hence 34 assignments with S greater than or equal to 110.5, so the exact one-sided p -value is $34/32,768 = 0.00104$.

Normal Approximation

This method of directly counting more extreme cases in order to find the p -value can be replaced by tables or by computer-assisted counting. A normal approximation for convenient computation of an approximate p -value is available. The mean and standard deviation of S in its permutation distribution are

$$\text{Mean}(S) = n(n + 1)/4 \quad \text{and} \quad \text{SD}(S) = [n(n + 1)(2n + 1)/24]^{1/2}.$$

Comparing Z -statistic = $[S - \text{Mean}(S)]/\text{SD}(S)$, with a continuity correction, to a standard normal distribution gives a good approximation to the p -value for $n \geq 20$.

For the schizophrenia study, $\text{Mean}(S) = 60$, $\text{SD}(S) = 17.61$, so that $Z\text{-stat} = 2.84$. An approximate one-sided p -value is 0.00226.

4.5 RELATED ISSUES

4.5.1 Practical and Statistical Significance

P -values indicate *statistical significance*, the extent to which a null hypothesis is contradicted by the data. This must be distinguished from *practical significance*, which describes the practical importance of the effect in question.

A study may find a statistically significant increase in rainfall due to cloud seeding, but the amount of increase in rainfall, say 10%, may not be sufficient to justify the expense of seeding. A finding that annual male salaries tend to be \$16.32 larger than female salaries, when annual incomes are on the order of \$10,000, does not provide a strong indication of persistent discrimination, even if the difference is statistically significant. A statistically significant reduction in time to complete a test problem may not be practically relevant if the reduction is only from 150 seconds to 147 seconds. These are issues of practical significance. They have little to do with statistics, but they are highly relevant to the interpretation of statistical analyses.

It is important to understand the connection between sample size and statistical significance. If there is a difference in population means (no matter how practically insignificant), large enough samples should indicate the existence of a statistically significant difference. On the other hand, even if there is a practically significant difference in population means (that is, an important difference), a small sample size may fail to indicate the existence of a statistically significant difference between means.

Three practical points deserve attention:

1. p -values are sample-size-dependent.
2. A result with a p -value of 0.08 can have more scientific relevance than one with a p -value of 0.001.

3. Tests of hypotheses by themselves rarely convey the full significance of the results. They should be accompanied by confidence intervals to indicate the range of likely effects and to facilitate the assessment of practical significance.

4.5.2 The Presentation of Statistical Findings

The Statistical Conclusions sections that accompany the case studies in this book are intended as models for communicating results. The following are some general recommendations.

1. State the conclusions as they apply to the question of interest and avoid statistical jargon and symbols, except to include appropriate measures of statistical uncertainty and to state which statistical tools were used.
2. Make clear exactly what is being estimated to answer a question of interest; for example, the difference in population means, a ratio of medians, or an additive treatment effect.
3. Prefer confidence intervals to standard errors, particularly when the estimate does not have a normal sampling (or randomization) distribution.
4. Use graphical displays to help convey important features of the data.
5. Do not include unnecessarily large numbers of digits in estimates or p -values.
6. If transformations have been used, express the results on the original scale of measurement.
7. Comment on the scope of inference. Was random assignment used? Was random sampling used?

4.5.3 Levene's (Median) Test for Equality of Two Variances

Sometimes a question of interest calls for a test of equality of two population variances. The *F-test for equal variances* and its associated confidence interval are available in standard statistical computer packages, but they are not robust against departures from normality. For example, p -values can easily be off by a factor of 10 if the distributions have shorter or longer tails than the normal.

A robust alternative is *Levene's test* (based on deviations from the median). Suppose there are n_1 observations Y_{1i} from population 1, and n_2 observations Y_{2i} from population 2. Let Z_{1i} be the absolute value of the deviation of the i th observation in group 1 from its group median: $|Y_{1i} - \text{median}_1|$, and let Z_{2i} be the absolute value of the deviation of the i th observation in group 2 from its median: $|Y_{2i} - \text{median}_2|$. The typical size of the Z 's indicates the degree of variability in each group. The Levene test idea is to perform a two-sample t -test on the Z 's to judge equal variability in the two groups. This procedure seems to have good power in detecting nonequal variability yet works well even for nonnormally distributed Y 's.

Example—Sex Discrimination

The data of Section 1.1.2 are used here to illustrate Levene's test for the hypothesis that the variance of male salaries is equal to the variance of female salaries. The test compares the 32 absolute deviations from group median of the males to the 61 absolute deviations for the females. The t -statistic is 0.4331 and the two-sided

p-value, by comparison to a *t*-distribution on 91 degrees of freedom, is 0.67. Thus Levene's test gives no evidence that the variances are unequal.

4.5.4 Survey Sampling

Survey sampling concerns selecting members of a specific (finite) population to be included in a survey and the subsequent analysis. One sampling method already discussed is *simple random sampling*, in which each subset of n items from the population has the same chance of being selected as the sample. This is appealing for its mathematical and conceptual simplicity, but is often practically unrealistic. The prohibitive part of finding a simple random sample of American voters, for example, is first finding a list of all American voters.

Instead of simple random sampling, survey organizations rely on complex sampling designs, which make use of stratification, multistage sampling, and cluster sampling. In *stratified sampling*, the population is divided into strata, based on supplementary information, and separate simple random samples are selected for each stratum. For example, samples of American voters may be taken separately for different states. *Multistage sampling*, as its name suggests, uses sampling at different stages. For example, a random sample of states is taken, then random samples of counties are taken within the selected states, and then random samples of individuals are selected from lists in the chosen counties. In *cluster sampling*, a simple random sample of clusters is selected at some stage of the sampling and all members within the cluster are included in the survey. For example, a random sample of households (the clusters) in a county may be selected and *all* voters in each household surveyed.

Standard Errors from Complex Sampling Designs

The formulas for standard errors so far presented are based on simple random sampling *with replacement*, which means that a member of a population can appear in a sample more than once. Random sampling in sample surveys, however, is conducted *without replacement* and does not permit a member to appear in a sample twice. Consequently, different standard errors must be calculated. The variance of an average from random sampling without replacement differs from that with replacement by a multiplicative factor called the *finite population correction* (FPC). If N is the population size and n the sample size, $\text{FPC} = (N - n)/N$. So the difference is small if only a small fraction of the population is sampled.

In addition, the usual standard errors are inappropriate for data from complex sampling designs. The ratio of the sampling variance of an estimate from data based on a particular sampling design to the sampling variance it would have had if it was based on a simple random sample of the same size is called the *design effect*. Design effects for estimates from complex stratified, multistage sampling designs are usually greater than 1. Therefore, application of standard statistical theory to complex sampling designs usually results in an overstatement of the degree of precision. For example, confidence intervals will be narrower than they really should be. Methods for obtaining correct standard errors are discussed in texts on survey sampling.

Nonresponse

Of 979 Vietnam veterans who were deemed to be eligible for the study in Section 3.1.2, only 900 were located, only 871 of those completed the first interview, only 665 of those were available for blood samples, and only 646 of those had valid dioxin readings. Assuming the 979 represented the population of interest, is there any danger in using the available 646 (66%) who responded and provided valid results to represent the population? The answer depends on whether those who were lost along the way differed in their dioxin levels from those who were not. The question that must be asked is whether or not being unavailable or not providing valid results had anything to do with dioxin level. If not, there is little harm in ignoring the nonresponders.

In sample surveys, the issue is usually much more serious, since those individuals who tend to respond to surveys often have much different views from those who do not. As an illustration, consider the popular telephone-in surveys that news programs run after presidential debates. Viewers call in (and have to pay for their calls) to indicate who they think won. Although this is entertaining, the results give an entirely unreliable indication of what all viewers think, since those who respond tend to be a quite different crowd from those who do not.

4.6 SUMMARY

The Space Shuttle O-Ring Study

The main feature of these data is that the two-sample t -tools cannot be relied on to provide valid measures of uncertainty. They are not sufficiently robust with such small, unbalanced samples of discrete data. The permutation test, however, can be applied, because it does not rely on any distributional assumptions. The computations are tedious, and inferences apply only to the 24 launches.

It should be mentioned that some information is lost due to the form in which the data are presented. They have been rather artificially divided into cool temperature launches and warm temperature launches, whereas the actual temperature data for each launch are available. The more advanced method of Poisson log-linear regression is applied to these data in Chapter 22.

The Cognitive Load Study

These data from a randomized experiment could be suitably analyzed with any of a number of techniques if it were not for the censoring of five observations. Without having actual completion times for these five students, it is impossible to take an average. The rank-sum test is convenient in this case since one only needs to know that these five students were tied for last. It can be used to test for an additive treatment effect and, with some effort, to provide a confidence interval for the treatment effect.

It should be mentioned that the rank-sum test is not suitable for censored observations if the rank of the censored observation cannot be determined. Thus, it

is limited to handling censoring for data problems like this one, in which a number of censored observations are tied for last.

4.7 EXERCISES

Conceptual Exercises

1. **Cognitive Load.** Suppose that there were two textbooks on coordinate geometry, one written with conventional worked problems and the other with modified worked problems. And suppose it is possible to identify a number of schools in which each of the textbooks is used. If you took random samples of size 14 from schools with each text and obtained exactly the same data as in the example in Section 4.1.2, would the analysis and conclusions be the same?
2. **O-Ring Data.** (a) Is it appropriate to use the two-sample *t*-test to compare the number of O-ring incidents in cold and warm launches? (b) Is it appropriate to use the rank-sum test? (c) Is it appropriate to use a permutation test based on the *t*-statistic?
3. **O-Ring Data.** When these data were analyzed prior to the *Challenger* disaster it was noticed that variability was greater in the group with the larger average, so a log transformation was used. Since the log of zero does not exist, all the zeros were deleted from the data set. Does this seem like a reasonable approach?
4. **O-Ring Data.** Explain why the two-sided *p*-value from the permutation test applied to the O-ring data is equal to the one-sided *p*-value (see Display 4.10).
5. **O-Ring Data.** If you looked at the source of the O-ring data and found that temperatures for each launch were recorded in degrees F rather than as over/under 65°F, what question would that raise? Would the answer affect your conclusions about the analysis?
6. **Motivation and Creativity.** In what way is the *p*-value for the motivation and creativity randomized experiment (Section 1.1.1) dependent on an assumed model?
7. Are there occasions when both the two-sample *t*-test and the rank-sum test are appropriate?
8. Can the rank-sum test be used for comparing populations with unequal variances?
9. Suppose that two drugs are both effective in prolonging length of life after a heart attack. Substantial statistical evidence indicates that the mean life length for those using drug A is greater than the mean life length for those using drug B, but the variation of life lengths for drug A is substantially greater as well. Explain why it is difficult to conclude that drug A is better even though the mean life length is greater.
10. In a certain problem, the randomization test produces an exact two-sided *p*-value of 0.053, while the *t*-distribution approximation produces 0.047. One might say that since the *p*-values are on opposite sides of 0.05, they lead to quite different conclusions and, therefore, the approximation is not adequate. Comment on this statement.
11. What is the difference between a permutation test and a randomization test?
12. Explain what is meant by the comment that there is no single test called a randomization test.
13. What confounding factors are possible in the O-ring failure problem?

Computational Exercises

14. **O-Ring Study.** Find the *t*-distribution approximation to the *p*-value associated with the observed *t*-statistic. Compare this approximation to the (correct) permutation test *p*-value.

- 15.** Consider these artificial data:

Group 1:	1	5	
Group 2:	4	8	9

The difference in averages $\bar{Y}_1 - \bar{Y}_2$ is -4 . What is a one-sided p -value from the permutation distribution of the *difference in averages*? (*Hint:* List the 10 possible groupings; compute the difference in averages for each of these groupings, then calculate the proportion of these less than or equal to -4 .)

- 16.** Consider these artificial data:

Group 1:	5	7	12
Group 2:	4	6	

Calculate a p -value for the hypothesis of no difference, using the permutation distribution of the difference in sample averages. (You do not need to calculate the t -statistic for each grouping, only the difference in averages.)

- 17. O-Ring Study.** Suppose the O-ring data had actually turned out as shown in Display 4.13. These are the same 24 numbers as before, but with the 2 and 3 switched. What is the one-sided p -value from the permutation test applied to the t -statistic? (This can be answered by examining Display 4.10.)

DISPLAY 4.13

Hypothetical O-ring data

Launch temperature	Number of O-ring incidents											
Below 65°F	1	1	1	2								
Above 65°F	0	0	0	0	0	0	0	0	0	0	0	1

- 18.** Suppose that six persons have an illness. Three are randomly chosen to receive an experimental treatment, and the remaining three serve as a control group. After treatment, a physician examines all subjects and assigns ranks to the severity of their symptoms. The patient with the most severe condition has rank 1, the next most severe has rank 2, and so on up to 6. It turns out that the patients in the treatment group have ranks 3, 5, and 6. The patients in the control group have ranks 1, 2, and 4. Is there any evidence that the treatment has an effect on the severity of symptoms? Use the randomization distribution of the sum of ranks in the treatment group to obtain a p -value. (First find the sum of ranks in the treatment group. Then write down all 20 groupings of the 6 ranks; calculate the sum of ranks in the treatment group for each. What proportion of these give a rank-sum as large as or larger than the observed one?)

- 19. Bumpus's Study.** Use a statistical computer program to perform the rank-sum comparison of humerus lengths in the sparrows that survived and the sparrows that perished (Exercise 2.21). (a) What is the two-sided p -value? (b) Does the statistical computer package report the exact p -value or the one based on the normal approximation? (c) If it reports the one using the normal approximation, does it use a continuity correction to the Z -statistic? (d) How does the p -value from the rank-sum test compare to the one from the two-sample t -test (0.08) and the one from the two-sample t -test when the smallest observation is set aside (see Chapter 3, Exercise 28)? (e) Explain the relative merits of (i) the two-sample t -test using the strategy for dealing with outliers and (ii) the rank-sum test.

- 20. Trauma and Metabolic Expenditure.** For the data in Exercise 18 in Chapter 3: (a) Determine the rank transformations for the data. (b) Calculate the rank-sum statistic by hand (taking the trauma

patients to be group 1.) (c) Mimic the procedures used in Display 4.5 and Display 4.7 to compute the Z -statistic. (d) Find the one-sided p -value as the proportion of a standard normal distribution larger than the observed Z -statistic.

21. Trauma and Metabolic Expenditure. Use a statistical computer package to verify the rank-sum and the Z -statistic obtained in Exercise 20. Is the p -value the same? (Does the statistical package use a continuity correction?)

22. Trauma and Metabolic Expenditure. Using the rank-sum procedure, find a 95% confidence interval for the difference in population medians: the median metabolic expenditure for the population of trauma patients minus the median metabolic expenditure for the population of nontrauma patients.

23. Motivation and Creativity. Use a statistical computer package to compute the randomization test's two-sided p -value for testing whether the treatment effect is zero for the data in Section 1.1.1 (file case0101). How does this compare to the results from the two-sample *t*-test (which is used as an approximation to the randomization test)?

24. Motivation and Creativity. Find a 95% confidence interval for the treatment effect (poem creativity score after intrinsic motivation questionnaire minus poem creativity score after extrinsic motivation questionnaire, from Section 1.1.1 (file case0101)) using the rank-sum procedure. (Use a statistical computer program.) How does this compare to the *t*-based confidence interval for the treatment effect?

25. Guinea Pig Lifetimes. Use the Welch *t*-tools to find a two-sided p -value and confidence interval for the effect of treatment on lifetimes of guinea pigs in Chapter 2, Exercise 11. Does the additive treatment effect seem like a sensible model for these data?

26. Schizophrenia Study. (a) Draw a histogram of the differences in hippocampus volumes in Display 4.12. Is there evidence that the population of differences is skewed? (b) Take the logarithms of the volumes for each of the 30 subjects, take the differences of the log volumes, and draw a histogram of these differences. Does it appear that the distribution of differences of log volumes is more nearly symmetric? (c) Carry out the paired *t*-test on the log-transformed volumes. How does the two-sided p -value compare with the one obtained on the untransformed data? (d) Find an estimate of and 95% confidence interval for the mean difference in log volumes. Back-transform these to get an estimate and confidence interval for the median of the population of ratios of volumes.

27. Schizophrenia Study. Find the two-sided p -value using the signed-rank test, as in Display 4.12, but after taking a log transformation of the hippocampus volumes. How does the p -value compare to the one from the untransformed data? Is it apparent from histograms that the assumptions behind the signed-rank test are more appropriate on one of these scales?

28. Darwin Data. Charles Darwin carried out an experiment to study whether seedlings from cross-fertilized plants tend to be superior to those from self-fertilized plants. He covered a number of plants with fine netting so that insects would be unable to fertilize them. He fertilized a number of flowers on each plant with their own pollen and he fertilized an equal number of flowers on the same plant with pollen from a distant plant. (He did not say how he decided which flowers received which treatments.) The seeds from the flowers were allowed to ripen and were set in wet sand to germinate. He placed two seedlings of the same age in a pot, one from a seed from a self-fertilized flower and one from a seed from a cross-fertilized flower. The data in Display 4.14 are the heights of the plants at certain points in time. (The fertilization experiments were described by Darwin in an 1878 book; these data were found in D. F. Andrews and A. M. Herzberg, *Data* (New York: Springer-Verlag, 1985), pp. 9–12.) (a) Draw a histogram of the differences. (b) Find a two-sided p -value for the hypothesis of no treatment effect, using the paired *t*-test. (c) Find a 95% confidence interval for the additive treatment effect. (d) Is there any indication from the plot in (a) that the paired

DISPLAY 4.14

Darwin's data: heights (inches) for 15 pairs of plants of the same age, one of which was grown from a seed from a cross-fertilized flower and the other of which was grown from a seed from a self-fertilized flower; first 5 of 15 rows

Plant height (inches)		
Pair	Cross-fertilized	Self-fertilized
1	23.5	17.375
2	12	20.375
3	21	20
4	22	20
5	19.125	18.375

t-test may be inappropriate? (e) Find a two-sided *p*-value for the hypothesis of no treatment effect for the data in Display 4.14, using the signed-rank test.

Data Problems

29. Salvage Logging. When wildfires ravage forests, the timber industry argues that logging the burned trees enhances forest recovery. The 2002 Biscuit Fire in southwest Oregon provided a test case. Researchers selected 16 fire-affected plots in 2004—before any logging was done—and counted tree seedlings along a randomly located transect pattern in each plot. They returned in 2005, after nine of the plots had been logged, and counted the tree seedlings along the same transects. (Data from D.C. Donato et al., 2006. “Post-Wildfire Logging Hinders Regeneration and Increases Fire Risk,” *Science*, 311: 352.) The numbers of seedlings in the logged (L) and unlogged (U) plots are shown in Display 4.15.

DISPLAY 4.15

Number of tree seedlings per transect in nine logged (L) and seven unlogged (U) plots affected by the Biscuit Fire, in 2004 and 2005, and the percentage of seedlings lost between 2004 and 2005; first 5 of 16 rows

Plot	Action	Seedlings2004	Seedlings2005	PercentLost
Plot 1	L	298	164	45.0
Plot 2	L	471	221	53.1
Plot 3	L	767	454	40.8
Plot 4	L	576	141	75.5
Plot 5	L	407	217	46.7

Analyze the data to see whether logging has any effect on the distribution of percentage of seedlings lost between 2004 and 2005, possibly using the following suggestions: (a) Use the rank-sum procedure to test for a difference between logged and unlogged plots. Also use the procedure to construct a 95% confidence interval on the difference in medians. (b) Use the *t*-tools to test for differences in mean percentages lost and to construct a 95% confidence interval. Compare the results with those in (a).

DISPLAY 4.16

Tolerance to sunlight (minutes) for 13 patients prior to treatment and after treatment with a sunscreen; first 5 of 13 rows

Tolerance to sunlight (minutes)		
Patient	Pretreatment	During treatment
1	30	120
2	45	240
3	180	480
4	15	150
5	200	480

DISPLAY 4.17

Months of survival after beginning of study for 58 breast cancer patients

Control Patients (*n* = 24)

2, 6, 8, 10, 12, 12, 14, 14, 14, 16, 16, 16, 16, 18, 18, 18, 18, 20, 22, 22, 26, 34, 36, 38, 40, 48

Patients Given Group Therapy for One Year (*n* = 34)

2, 2, 4, 4, 4, 6, 6, 8, 10, 10, 12, 14, 16, 16, 16, 18, 20, 22, 32, 36, 46, 46, 48, 48, 58, 58, 66, 72, 72, 82, 122, 122*, 122*, 122*

*These three patients were still alive at the end of the 122-month study period.

30. Sunlight Protection Factor. A sunscreen sunlight protection factor (SPF) of 5 means that a person who can tolerate Y minutes of sunlight without the sunscreen can tolerate $5Y$ minutes of sunlight with the sunscreen. The data in Display 4.16 are the times in minutes that 13 patients could tolerate the sun (a) before receiving treatment and (b) after receiving a particular sunscreen treatment. (Data from R. M. Fusaro and J. A. Johnson, "Sunlight Protection for Erythropoietic Protoporphyrin Patients," *Journal of the American Medical Association* 229(11) (1974): 1420.) Analyze the data to estimate and provide a confidence interval for the sunlight protection factor. Comment on whether there are any obvious potentially confounding variables in this study.

31. Effect of Group Therapy on Survival of Breast Cancer Patients. Researchers randomly assigned metastatic breast cancer patients to either a control group or a group that received weekly 90-minute sessions of group therapy and self-hypnosis, to see whether the latter treatment improved the patients' quality of life. The group therapy involved discussion and support for coping with the disease, but the patients were not led to believe that the therapy would affect the progression of their disease. Surprisingly, it was noticed in a follow-up 10 years later that the group therapy patients appeared to have lived longer. The data on number of months of survival after beginning of the study are shown in Display 4.17. (Data from a graph in D. Spiegel, J. R. Bloom, H. C. Kraemer, and E. Gottheil, "Effect of Psychosocial Treatment on Survival of Patients with Metastatic Breast Cancer," *Lancet* (October 14, 1989): 888–91.) Notice that three of the women in the treatment group were still alive at the time of the follow-up, so their survival times are only known to be larger than 122 months. Is there indeed evidence of an effect of the group therapy treatment on survival time and, if so, how much more time can a breast cancer patient expect to live if she receives this therapy? Analyze the data as best as possible and write a brief report of the findings.

32. Therapeutic Marijuana. Nausea and vomiting are frequent side effects of cancer chemotherapy, which can contribute to the decreased ability of patients to undergo long-term chemotherapy

DISPLAY 4.18

Number of vomiting and retching episodes for 15 chemotherapy-receiving cancer patients, under placebo and marijuana treatments; first 5 of 15 rows

Subject number	Total number of vomiting and retching episodes	
	Marijuana	Placebo
1	15	23
2	25	50
3	0	0
4	0	99
5	4	31

schedules. To investigate the capacity of marijuana to reduce these side effects, researchers performed a double-blind, randomized, crossover trial. Fifteen cancer patients on chemotherapy schedules were randomly assigned to receive either a marijuana treatment or a placebo treatment after their first three chemotherapy sessions, and then “crossed over” to the opposite treatment after their next three sessions. The treatments, which involved both cigarettes and pills, were made to appear the same whether in active or placebo form. Shown in Display 4.18 are the number of vomiting and retching episodes for the 15 subjects. Does marijuana treatment reduce the frequency of episodes? By how much? Analyze the data and write a statistical summary of conclusions. (Data from A. E. Chang et al., “Delta-9-Tetrahydrocannabinol as an Antiemetic in Cancer Patients Receiving High-Dose Methotrexate,” *Annals of Internal Medicine*, Dec. 1979. The order of the treatments is unavailable.)

Answers to Conceptual Exercises

1. The analysis would be the same. The conclusions would be very different. You could infer a real difference in the solution times of the two groups, but you could not attribute it to the different text types because of the possibility of a host of confounding factors.
2. (a) No, the extent of the nonnormality in these small and unequally sized samples is more than can be tolerated by the two-sample t -test. (b) Probably not. The spreads apparently are not equal. (c) Yes. The permutation test for significance requires no model or assumptions.
3. No! Observations cannot be deleted simply because the transformation does not work on them. In this case, a major portion of the data was deleted, leaving a very misleading picture.
4. There are 105 groupings that lead to t -statistics greater than or equal to 3.888 and no groupings that lead to t -statistics less than or equal to -3.888 .
5. Was the 65°F cutoff chosen to maximize the apparent difference in the two groups? If so, the p -value would not be correct. Why? Because the p -value represents the chance of getting evidence as damaging to the hypothesis when there is no difference. The “chance” is the frequency of occurrence in replications of the study, *using the same statistical analysis*. The p -value calculation assumes that the 65°F cutoff will always define the groups. If the choice of cutoff was part of the statistical analysis used on this data set, the calculation was not correct. To get a correct p -value would require that you allow for a different cutoff to be chosen in each replication. Further discussion of data snooping is given in Chapter 6.
6. The p -value is based on the two-sample t -test but it is now understood that this p -value serves as an approximation to the p -value from the exact randomization test. For this approximation to be valid, the histograms of creativity scores should be reasonably normal (which they are).

7. Yes. Since the population model for the *t*-tools requires that the populations be normal with equal spread they will necessarily have the same shape and spread. Therefore, the assumptions for the rank-sum test are also satisfied.
8. Yes, but the meaning may be unclear if the variances are substantially unequal.
9. Generally, it is hard to make use of the difference in the centers of two distributions when the spreads are quite different. Specifically, the mean life length for drug A may be longer, but more people who use it may die sooner than for drug B.
10. The statement takes the rejection region approach too literally. There is very little difference in the degree of evidence against the null hypothesis in *p*-values of 0.053 and 0.047, so the approximation is pretty good.
11. A randomization test is a permutation test applied to data from a randomized experiment.
12. There is a different permutation distribution for each statistic that can be calculated from the data.
13. Perhaps workers tended to make more mistakes in cold weather or wind stress was greater on days with cold weather.

Comparisons Among Several Samples

The issues and tools associated with the analysis of three or more independent samples (or treatment groups in a randomized experiment) are very similar to those for comparing two samples. Notable differences, however, stem from the particular kinds of questions and the greater number of them that may be asked.

An initial question, often asked in a preliminary stage of the analysis, is whether all of the means are equal. An easy-to-use F -test is available for investigating this. A typical analysis, however, goes beyond the F -test and explores linear combinations of the means to address particular questions of interest. A simple example of a linear combination of means is $\mu_3 - \mu_1$, the difference between means in populations 3 and 1. Inferences about this parameter can be made with t -tools just as for the two-independent-sample problem, with the important difference that the pooled estimate of standard deviation is from all groups, not just from those being compared.

This chapter discusses the use of the pooled estimate of variance for specific linear combinations of means and the one-way analysis of variance F -test for testing the equality of several means. The next chapter looks at linear combinations more generally and the problem of compounded uncertainty from the multiple, simultaneous comparisons of means.

5.1 CASE STUDIES

5.1.1 Diet Restriction and Longevity—A Randomized Experiment

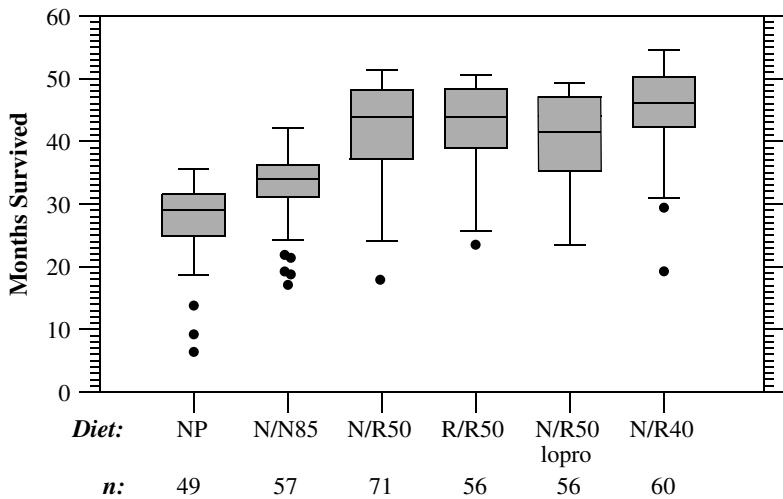
A series of studies involving several species of animals found that restricting caloric intake can dramatically increase life expectancy. In one such study, female mice were randomly assigned to one of the following six treatment groups:

- NP:** Mice in this group ate as much as they pleased of a nonpurified, standard diet for laboratory mice.
- N/N85:** This group was fed normally both before and after weaning. (The slash distinguishes the two periods.) After weaning, the ration was controlled at 85 kcal/wk. This, rather than NP, serves as the control group because caloric intake is held reasonably constant.
- N/R50:** This group was fed a normal diet before weaning and a reduced-calorie diet of 50 kcal/wk after weaning.
- R/R50:** This group was fed a reduced-calorie diet of 50 kcal/wk both before and after weaning.
- N/R50 lopro:** This group was fed a normal diet before weaning, a restricted diet of 50 kcal/wk after weaning, and had dietary protein content decreased with advancing age.
- N/R40:** This group was fed normally before weaning and was given a severely reduced diet of 40 kcal/wk after weaning.

Display 5.1 shows side-by-side box plots for the lifetimes, measured in months, of the mice in the six groups. Summary statistics and sample sizes are reported in

DISPLAY 5.1

Lifetimes of female mice fed on six different diet regimens



DISPLAY 5.2

Summary statistics for lifetimes of mice on six different diet regimens

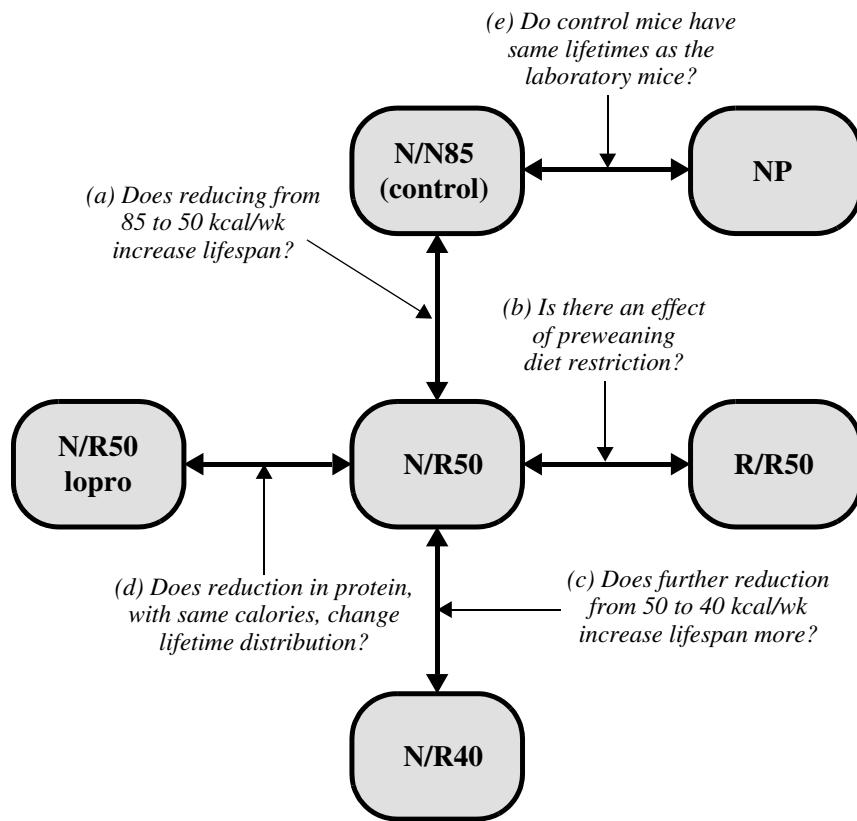
Group	n	Range (months)	Average	SD	95% CI for mean
NP	49	6.4–35.5	27.4	6.1	25.6–29.2
N/N 85	57	17.9–42.3	32.7	5.1	31.3–34.1
N/R50	71	18.6–51.9	42.3	7.8	40.5–44.1
R/R50	56	24.2–50.7	42.9	6.7	41.1–44.7
N/R50 lopro	56	23.4–49.7	39.7	7.0	37.8–41.6
N/R40	60	19.6–54.6	45.1	6.7	43.4–46.8

Display 5.2. (Data from R. Weindruch, R. L. Walford, S. Fligiel, and D. Guthrie, “The Retardation of Aging in Mice by Dietary Restriction: Longevity, Cancer, Immunity and Lifetime Energy Intake,” *Journal of Nutrition* 116(4) (1986): 641–54.)

The questions of interest involve specific comparisons of treatments as diagrammed in Display 5.3. Specifically, (a) Does lifetime on the 50 kcal/wk diet exceed

DISPLAY 5.3

Structure of planned comparisons among groups in the diet restriction study



the lifetime on the 85 kcal/wk diet? If so, by how much? (This calls for a comparison of the N/R50 group to the N/N85 group.) (b) Is lifetime affected by providing a reduced calorie diet before weaning, given that a 50 kcal/wk diet is provided after weaning? (This calls for a comparison of the R/R50 group to the N/R50 group.) (c) Does lifetime on the 40 kcal/wk diet exceed the lifetime on the 50 kcal/wk diet? (This calls for a comparison of the N/R40 group to the N/R50 group.) (d) Given a reduced calorie diet of 50 kcal/week, is there any additional effect on lifetime due to decreasing the protein intake? (This calls for a comparison of the N/R50 lopro diet to the N/R50 diet.) (e) Is there an effect on lifetime due to restriction at 85 kcal/week? This would indicate the extent to which the 85 kcal/wk diet served as a proper control and possibly whether there was any effect of restricting the diet, even with a standard caloric intake. (This calls for a comparison of the N/N85 group to the NP group.) Comparisons other than those indicated by the arrows are not directly meaningful because the group treatments differ in more than one way. The N/R50 lopro and the N/R40 groups, for example, differ in both the protein composition and the total calories in the diet, so a difference would be difficult to attribute to a single cause.

Statistical Conclusion

These data provide overwhelming evidence that mean lifetimes in the six groups are different (p -value < 0.0001 ; analysis of variance F -test). Answers to the five particular questions are indicated as follows:

- (a) There is convincing evidence that lifetime is increased as a result of restricting the diet from 85 kcal/wk to 50 kcal/wk (one-sided p -value < 0.0001 ; t -test). The increase is estimated to be 9.6 months (95% confidence interval: 7.3 to 11.9 months).
- (b) There is no evidence that reducing the calories before weaning increased lifetime, when the calorie intake after weaning is 50 kcal/wk (one-sided p -value = 0.32; t -test). A 95% confidence interval for the amount by which the lifetime under the R/R50 diet exceeds the lifetime under the N/R50 diet is -1.7 to 2.9 months.
- (c) Further restriction of the diet from 50 to 40 kcal/wk increases lifetime by an estimated 2.8 months (95% confidence interval: 0.5 to 5.1 months). The evidence that this effect is greater than zero is moderate (one-sided p -value = 0.017; t -test). (The combined effect of the reduction from 85 to 40 kcal/wk is estimated to be 12.4 months. This is a 38% increase in mean lifetime. If extended to human subjects, a 50% reduction in caloric intake might increase typical lifetimes—of 80 years—to 110 years.)
- (d) There was moderate evidence that lifetime was *decreased* by the lowering of protein in addition to the 50 kcal/wk diet (two-sided p -value = 0.024; t -test). The estimated decrease in lifetime is 2.6 months (95% confidence interval: 0.3 to 4.9 months).
- (e) There is convincing evidence that the control mice live longer than the mice on the nonpurified diet (one-sided p -value < 0.0001).

5.1.2 The Spock Conspiracy Trial—An Observational Study

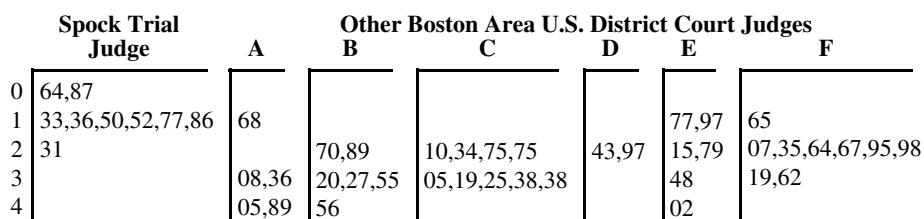
In 1968 Dr. Benjamin Spock was tried in United States District Court of Massachusetts in Boston on charges of conspiring to violate the Selective Service Act by encouraging young men to resist being drafted into military service for Vietnam. The defense in that case challenged the method by which jurors were selected, claiming that women—many of whom had raised children according to popular methods developed by Dr. Spock—were underrepresented. In fact, the Spock jury had no women.

Boston area juries are selected in three stages. From the City Directory, the Clerk of the Court selects at random 300 names for potential jury duty. Before a trial, a *venire* of 30 or more jurors is selected from the 300 names, again—according to the law—at random. An actual jury is selected from the venire in a nonrandom process allowing each side to exclude certain jurors for a variety of reasons.

The Spock defense pointed to the venire for their trial, which contained only one woman. That woman was released by the prosecution, making an all-male jury. Defense argued that the judge in the trial had a history of venires in which women were systematically underrepresented, contrary to the law. They compared this district judge's recent venires with the venires of six other Boston area district judges. The percents of women in those venires are presented in Display 5.4 as stem-and-leaf diagrams. (Data from H. Zeisel and H. Kalven, Jr., "Parking Tickets, and Missing Women: Statistics and the Law," in J. M. Tanur, F. Mosteller, W. H. Kruskal, R. F. Link, R. S. Pieters, and G. R. Rising, eds., *Statistics: A Guide to the Unknown*, San Francisco: Holden-Day, 1972.)

DISPLAY 5.4

Percents of women in 30-juror venires for Boston area U.S. District Court trials, grouped according to the judge presiding



Legend: 4|89 represents a venire with 48.9% women.

There are two key questions: (1) Is there evidence that women are underrepresented on the Spock judge's venires compared to the venires of the other judges? and (2) Is there any evidence that there are differences in women's representation in the venires of the other six judges? The first question addresses the key issue, but the second question has considerable bearing on the interpretation of the first. If the other judges all had about the same percentage of women on their venires while the Spock judge had significantly fewer women, this would make a strong statement

about that particular judge. But if the percentages of women in the venires of the other judges are all different, this would put a very different perspective on any difference that the Spock judge's venires would have from the average of the other judges' venires.

Statistical Conclusion

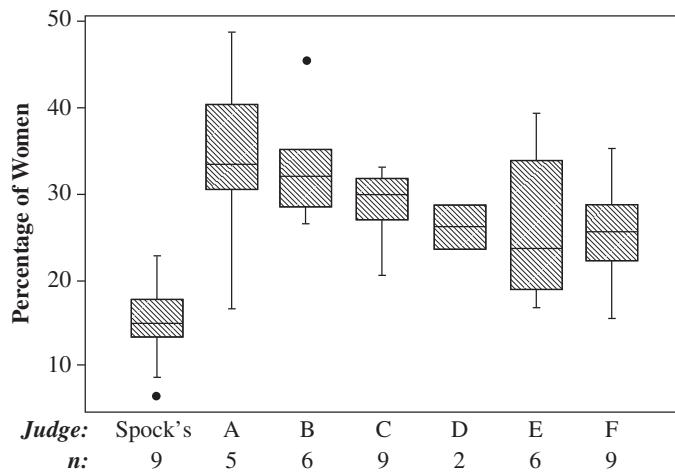
As evident from the box plots in Display 5.5, the percentages of women on Spock's judge's venires (with an average of 15%) were substantially lower than those of the other judges (with an average of 30%). The one-sided p -value from a two-sample t -test comparing the mean percentage of Spock's judge to the mean percentage of all others combined is less than 0.000001. This choice of comparison, which combines the venires from the six other judges into a single group, is supported by a lack of significant differences among the other judges (p -value = 0.32 from a one-way analysis of variance F -test of whether the mean percentage is the same for judges A–F). It is estimated that the mean percentage of women on Spock judge's venires is 15% less than the mean of other venires (with a 95% confidence interval of 10% to 20%). (Note: A separate approach that does not combine venires from judges A–F into a single group is detailed in Section 6.2.3, and yields essentially the same conclusion.)

Scope of Inference

There is no indication that the venires in this observational study were randomly selected, so inference to some population is speculative. Thinking of the p -values as approximate p -values for permutation tests, however, leads one to conclude that the Spock judge did have a lower proportion of females on his venires than did the other judges—more so than can be explained by chance.

DISPLAY 5.5

Percentages of women on venires of the seven Boston area judges



5.2 COMPARING ANY TWO OF THE SEVERAL MEANS

When subjects in a study are divided into distinct experimental or observational categories, the study itself is said to be a *several-group* problem, or a *one-way classification* problem. Mice were divided into six experimental groups; samples of venires were obtained for seven judges. A typical analysis of several-group data involves graphical exploration (like side-by-side box plots), consideration of transformations, initial screening to evaluate differences among all groups, and inferential techniques chosen to specifically address the questions of interest.

In the two-sample problem the questions of interest require inference about $\mu_2 - \mu_1$. In the several-sample problem the questions of interest may involve a few pairwise differences of means, like $\mu_2 - \mu_1$ and $\mu_3 - \mu_1$; all possible pairwise differences of means; or specific linear combinations of means, like

$$[-1 \times \mu_1] + [(1/2) \times \mu_2] + [(1/2) \times \mu_3].$$

This aspect of the data structure requires careful attention to how the questions of interest can be addressed through model parameters. If there are multiple questions, as, for example, “Does group 1 differ from group 2, does group 1 differ from group 3, and does group 2 differ from group 3?” then attention to interpretations of multiple, simultaneous statistical inferences is important. This *multiple comparison* problem is discussed further in the next chapter. For now, the discussion focuses on the single comparison of any two means.

5.2.1 An Ideal Model for Several-Sample Comparisons

The ideal population model, upon which the standard tools are derived, is a straightforward extension of the normal model for two-sample comparisons: (1) The populations have normal distributions. (2) The population standard deviations are all the same. (3) Observations within each sample are independent of each other. (4) Observations in any one sample are independent of observations in other samples.

As in the two-sample problem, the equal standard deviation model is entertained not because all data sets with several groups of measurements necessarily fit the description but because (1) it is conceptually difficult to compare populations with unequal variability, (2) it is statistically difficult as well, (3) for many problems a treatment is associated with an effect on the mean but not on the standard deviation, and (4) for many problems with unequal standard deviations it is possible to transform the data in such a way that the equal-spread model is valid on the transformed scale.

Notation

The symbols for population parameters are the same as before. A population mean is denoted by the Greek letter μ with a subscript indicating its group. The Greek letter σ is used to represent the standard deviation common to all the sampled populations. The number of samples will be represented by I , and the number of

observations within the i th sample will be represented by n_i . The total number of observations from all groups combined will be $n = n_1 + n_2 + \dots + n_I$. There are $I + 1$ parameters in the ideal model—the I means $\mu_1, \mu_2, \dots, \mu_I$, and the single standard deviation σ .

Treatment Effects for Randomized Experiments

The discussion that follows in this chapter will continue to use the terminology of samples rather than of treatment groups, even though the methods also apply to data from randomized experiments with I treatment groups. An additive treatment effect model asserts that an experimental unit that would produce a response of Y_1 on treatment 1 would produce a response of $Y_1 + \delta_1$ on treatment 2, $Y_1 + \delta_2$ on treatment 3, and so on. As before, exact randomization tests are available for inferences about the δ 's, but approximations based on the tools developed for random samples from populations are usually adequate. The practical upshot is that data from randomized experiments will be analyzed in exactly the same way as samples from populations, but concluding statements will be worded in terms of treatment effects rather than differences in population means.

5.2.2 The Pooled Estimate of the Standard Deviation

The mean from the i th population, μ_i , is estimated by the average from the i th sample, \bar{Y}_i . The variance σ^2 is estimated separately by s_i^2 from each of the I samples. If all the sample sizes are equal, then the best single estimate of σ^2 is their average. When the samples have different numbers of observations, a *weighted average* is more appropriate, where the estimates from larger samples are given more weight than those from smaller samples. The *pooled estimate of variance*, s_p^2 , is a weighted average of sample variances in which each sample variance receives the weight of its degrees of freedom:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \dots + (n_I - 1)s_I^2}{(n_1 - 1) + (n_2 - 1) + \dots + (n_I - 1)}.$$

Its associated degrees of freedom are the sum of degrees of freedom from all samples, $n - I$. An illustration of the calculations for the diet restriction data is shown in Display 5.6.

5.2.3 t-Tests and Confidence Intervals for Differences of Means

If \bar{Y}_i is the average based on a sample from the population with mean μ_i and variance σ_i^2 , and if \bar{Y}_j is the average based on an independent sample from the population with mean μ_j and variance σ_j^2 , then the sampling distribution of $\bar{Y}_i - \bar{Y}_j$ has variance $\sigma_i^2/n_i + \sigma_j^2/n_j$. If the two population variances are equal, this reduces to $\sigma^2[(1/n_i) + (1/n_j)]$. This may be estimated by replacing the unknown σ^2 by

DISPLAY 5.6

Pooled estimate of standard deviation; diet restriction data

Group	<i>n</i>	Sample SD
NP	49	6.1
N/N 85	57	5.1
N/R50	71	7.8
R/R50	56	6.7
N/R50 lopro	56	7.0
N/R40	60	6.7

$$s_p^2 = \frac{(49-1)(6.1)^2 + (57-1)(5.1)^2 + (71-1)(7.8)^2 + (56-1)(6.7)^2 + (56-1)(7.0)^2 + (60-1)(6.7)^2}{(49-1) + (57-1) + (71-1) + (56-1) + (56-1) + (60-1)}$$

$$= \frac{15,313.90}{343} = 44.647; \quad s_p = \sqrt{44.647} = 6.68$$

Calculate the pooled estimate of variance, s_p^2

s_p is the square root.

$d.f.$ is the denominator.

its best estimate. The important aspect of this in the one-way classification is that if the variances from all I populations can be assumed to be equal, then the best estimate is the pooled estimate of variance from all groups. So, for example, the standard error of $\bar{Y}_3 - \bar{Y}_2$ is

$$\text{SE}(\bar{Y}_3 - \bar{Y}_2) = s_p \sqrt{\frac{1}{n_3} + \frac{1}{n_2}},$$

where s_p is the pooled estimate from *all* groups, with $(n - I)$ degrees of freedom.

The theory leading to confidence intervals and tests is the same as for the two-independent-sample problem. In this case the *t*-ratio has a *t* distribution on $n - I$ degrees of freedom. A 95% confidence interval for $\mu_3 - \mu_2$ is $(\bar{Y}_3 - \bar{Y}_2) \pm t_{n-I}(0.975) \times \text{SE}(\bar{Y}_3 - \bar{Y}_2)$ and a *t*-statistic for testing the hypothesis that $\mu_3 - \mu_2$ equals zero is $(\bar{Y}_3 - \bar{Y}_2)/\text{SE}(\bar{Y}_3 - \bar{Y}_2)$. These are illustrated in Display 5.7.

5.3 THE ONE-WAY ANALYSIS OF VARIANCE *F*-TEST

One question often asked, possibly in a first stage of the analysis, is “are there differences between *any* of the means?” The *analysis of variance* (ANOVA) *F*-test provides evidence in answer to this question. The term *variance* should not mislead; this is most definitely a test about *means*. It assesses mean differences by comparing the amounts of variability explained by different sources.

DISPLAY 5.7

A confidence interval for $\mu_3 - \mu_2$ and a test that $\mu_3 - \mu_2 = 0$ (diet restriction data)

(1)

Get averages, sample sizes, and pooled estimate of standard deviation.

Group	3: N/R50	2: N/N85
Sample size	71	57
Average (months)	42.3	32.7

Pooled estimate of σ : $s_p = 6.68$ months; d.f. = 343 (from Display 5.6)

(2)

Compute the estimate of $\mu_3 - \mu_2$ and its standard error.

$$\text{Estimate: } \bar{Y}_3 - \bar{Y}_2 = 42.3 - 32.7 = 9.6 \text{ months}$$

$$\text{SE}(\bar{Y}_3 - \bar{Y}_2) = 6.68 \sqrt{\frac{1}{71} + \frac{1}{57}} = 1.2 \text{ months}$$

(3)

95% confidence interval for $\mu_3 - \mu_2$.

$$t_{343}(0.975) = 1.96$$

$$95\% \text{ CI: } 9.6 \pm (1.96)(1.2) = \boxed{7.3} \text{ months} \quad \boxed{11.9} \text{ months}$$

(4)

Test the hypothesis that $\mu_3 - \mu_2 = 0$.

$$t\text{-stat} = \frac{9.6}{1.2} = 8.08 \longrightarrow \text{two-sided } p\text{-value} < 0.0001$$

5.3.1 The Extra-Sum-of-Squares Principle

One model for the Spock data is that the percentage of women on venires for judge i comes from a normal distribution with mean μ_i and variance σ^2 , for i from 1 to 7. A hypothesis for initial exploration is that all seven means are equal. That is, the null hypothesis is $H: \mu_1 = \mu_2 = \dots = \mu_7$, and the alternative is that at least one of the means differs from the others.

The term *extra-sum-of-squares* refers to a general idea for hypothesis testing and is fundamental to a large class of tests. This section introduces the general notion of extra-sum-of-squares and the associated F -test, but with particular attention to testing the above hypothesis in the one-way classification problem.

Full and Reduced Models

Testing a hypothesis is formulated as a problem of comparing two models for the response means. A *full model* is a general model that is found to adequately describe

the data. The *reduced model* is a special case of the full model obtained by imposing the restrictions of the null hypothesis.

For comparing equality of all means in the several-sample problem, the full model is the one that has a separate mean for each group. The reduced model, obtained from the full model by supposing the hypothesis of equal means is true, specifies a single mean for all populations. For the Spock data, the means from the two models are the following:

Group:	1	2	3	4	5	6	7
Full (separate-means) model:	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7
Reduced (equal-means) model:	μ						

The terminology of *full* and *reduced* models provides a framework for the general procedure called the *extra-sum-of-squares F-test*. For any particular application, the models will have different names that refer more specifically to the problem at hand. As a test of equality of means in one-way classification, it makes more sense to call the full model the *separate-means* model and the reduced model the *equal-means* model.

Fitting the Models

The idea behind analysis of variance is to estimate the parameters in both the full and reduced models and to see whether the variability of responses about the estimated means is comparable in the two models. The *estimated* means for each group are different for the two models:

Group:	1	2	3	4	5	6	7
Full (separate-means) model:	\bar{Y}_1	\bar{Y}_2	\bar{Y}_3	\bar{Y}_4	\bar{Y}_5	\bar{Y}_6	\bar{Y}_7
Reduced (equal-means) model:	\bar{Y}						

where \bar{Y} is the average of all observations, called the *grand average*.

Residuals

Associated with each observation in the data set is an estimated group mean based on the full model and a different estimated group mean based on the reduced model. Also associated with each observation is a residual for each model. A residual is the observation value minus its estimated mean. So, for observation Y_{ij} (the percentage of women on the j th venire for judge i), the residual from the full model is $Y_{ij} - \bar{Y}_i$ and the residual from the reduced model is $Y_{ij} - \bar{Y}$. Display 5.8 shows the sets of estimated means and residuals for the Spock trial data in both the full (separate-means) and the reduced (equal-means) models. Notice that the residuals tend to be larger for the equal-means model.

If the null hypothesis is correct, then the two models should be about equal in their ability to explain the data, and the magnitudes of the residuals should be about the same. If the null hypothesis is incorrect, the magnitudes of the residuals from the equal-means model will tend to be larger.

DISPLAY 5.8

Estimated means and residuals from two models for mean percentage of women (%W) in venues: Spock trial data

Large residuals indicate that the model fits poorly.

Judge	%W	Equal means		Separate means		Judge	%W	Equal means		Separate means	
		Est.	Res.	Est.	Res.			Est.	Res.	Est.	Res.
Spock	6.4	26.6	-20.2	14.6	-8.2	C	21.0	26.6	-5.6	29.1	-8.1
Spock	8.7	26.6	-17.9	14.6	-5.9	C	23.4	26.6	-3.2	29.1	-5.7
Spock	13.3	26.6	-13.3	14.6	-1.3	C	27.5	26.6	0.9	29.1	-1.6
Spock	13.6	26.6	-13.0	14.6	-1.0	C	27.5	26.6	0.9	29.1	-1.6
Spock	15.0	26.6	-11.6	14.6	0.4	C	30.5	26.6	3.9	29.1	1.4
Spock	15.2	26.6	-11.4	14.6	0.6	C	31.9	26.6	5.3	29.1	2.8
Spock	17.7	26.6	-8.9	14.6	3.1	C	32.5	26.6	5.9	29.1	3.4
Spock	18.6	26.6	-8.0	14.6	4.0	C	33.8	26.6	7.2	29.1	4.7
Spock	23.1	26.6	-3.5	14.6	8.5	C	33.8	26.6	7.2	29.1	4.7
A	16.8	26.6	-9.8	34.1	-17.3	D	24.3	26.6	-2.3	27.0	-2.7
A	30.8	26.6	4.2	34.1	-3.3	D	29.7	26.6	3.1	27.0	2.7
A	33.6	26.6	7.0	34.1	-0.5	E	17.7	26.6	-8.9	27.0	-9.3
A	40.5	26.6	13.9	34.1	6.4	E	19.7	26.6	-6.9	27.0	-7.3
A	48.9	26.6	22.3	34.1	14.8	E	21.5	26.6	-5.1	27.0	-5.5
B	27.0	26.6	0.4	33.6	-6.6	E	27.9	26.6	1.3	27.0	0.9
B	28.9	26.6	2.3	33.6	-4.7	E	34.8	26.6	8.2	27.0	7.8
B	32.0	26.6	5.4	33.6	-1.6	E	40.2	26.6	13.6	27.0	13.2
B	32.7	26.6	6.1	33.6	-0.9	F	16.5	26.6	-10.1	26.8	-10.3
B	35.5	26.6	8.9	33.6	1.9	F	20.7	26.6	-5.9	26.8	-6.1
B	45.6	26.6	19.0	33.6	12.0	F	23.5	26.6	-3.1	26.8	-3.3
						F	26.4	26.6	-0.2	26.8	-0.4
						F	26.7	26.6	0.1	26.8	-0.1
						F	29.5	26.6	2.9	26.8	2.8
						F	29.8	26.6	3.2	26.8	3.0
						F	31.9	26.6	5.3	26.8	5.1
						F	36.2	26.6	9.6	26.8	9.4

In the equal-means model, estimated means are equal to the grand average.

In the separate-means model, estimated means are the group averages.

Residual Sums of Squares

A single summary of the magnitude of the residuals for a particular model is the *residual sum of squares* for that model (i.e., the sum of the squared residuals). By adding the squares of the two sets of residuals in Display 5.8 separately, one finds that the residual sum of squares for the equal-means model is 3,791.53 and the residual sum of squares from the separate-means model is 1,864.45. Is this large difference ($3,791.53 - 1,864.45 = 1,927.08$) due to the relatively poor fit of the equal-means model, or can it be explained by sampling variability? The *F*-test answers this question precisely.

General Form of the Extra-Sum-of-Squares *F*-Statistic

The *extra sum of squares* is the single number that summarizes the difference in sizes of residuals from the full and reduced models. As just calculated above,

$$\boxed{\begin{aligned} \text{Extra sum of squares} = \\ \text{Residual sum of squares (reduced)} - \text{Residual sum of squares (full)}. \end{aligned}}$$

A residual sum of squares measures the variability in the observations that remains unexplained by a model, so the extra sum of squares measures the amount of unexplained variability in the reduced model that is explained by the full model.

The extra sum of squares is a test statistic for judging the plausibility of the null hypothesis. For practical use, however, it must be converted into a form that can be compared to a known probability distribution. The particular form is the *F*-statistic:

$$\boxed{F\text{-statistic} = \frac{(\text{Extra sum of squares}) / (\text{Extra degrees of freedom})}{\hat{\sigma}_{\text{full}}^2}},$$

where the *extra degrees of freedom* are the number of parameters in the mean for the full model minus the number of parameters in the mean for the reduced model, and $\hat{\sigma}_{\text{full}}^2$ is the estimate of σ^2 based on the full model. Thus, the *F*-statistic is the extra sum of squares per extra degree of freedom, scaled by the best estimate of variance.

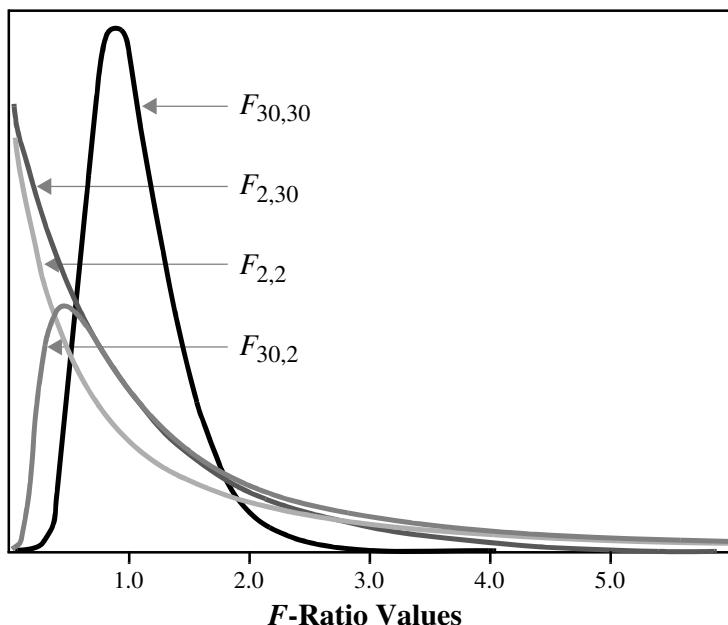
The *F*-Test

Large *F*-statistics are associated with large differences in the size of residuals from the two models. They supply evidence against the hypothesis of equal means and in favor of a model with different means. The test is summarized by its *p*-value, the chance of finding an *F*-statistic as large as or larger than the observed one when all means are, in fact, equal.

***F*-Distributions**

If all means are equal, the sampling distribution of the *F*-statistic is that of an *F-distribution*, which depends on two known parameters: the *numerator degrees of freedom* and the *denominator degrees of freedom*. For each degrees of freedom pair, there is a separate *F*-distribution. The letter *F* was given to this class of distributions by George Snedecor, honoring the British statistician Sir Ronald Fisher.

Theoretical histograms for four of the *F*-distributions are shown in Display 5.9. Notice that *F*-values in the general range of 0.5 to 3.0 are fairly typical. An *F*-statistic in this range would not be considered as very strong evidence of unequal means for most degree of freedom combinations. An *F*-statistic in the range from

DISPLAY 5.9Four F -distributions, having different degrees of freedom

3.0 to 4.0 would be highly unlikely with large degrees of freedom in both numerator and denominator but would be only moderately suggestive of differences with smaller degrees of freedom. An F -statistic larger than 4.0 is strong evidence against equal means except for the smallest degrees of freedom, particularly in the denominator. (Note: All the curves have the same area below them. This means that the $F_{2,2}$ and $F_{30,2}$ curves must have considerable area, spread thinly, off the graph to the right.)

Statistical computer programs and some calculators can provide tail area probabilities of F -distributions for specified F -values and, conversely, can provide the F -values corresponding to specified tail area probabilities. In both cases, the user must also specify the particular F -distribution of interest by supplying the numerator and denominator degrees of freedom. The p -value indicated by bubble 7 in Display 5.10, showing the F -test results for the Spock trial data, was obtained with a statistical computer package.

5.3.2 The Analysis of Variance Table for One-Way Classification

The analysis of variance table organizes and displays the calculations used in the F -test. Analysis of variance tables extend to more complicated structures, where several F -tests simultaneously evaluate how well different models fit the data. The ANOVA table for the Spock data appears in Display 5.10.

DISPLAY 5.10

Analysis of variance table: a test for equal mean percentages of women in venires of seven judges (Spock data)

Source of Variation	Sum of Squares	d.f.	Mean Square	<i>F</i> -Statistic	p-Value
Between Groups	1,927.08	6	321.18	6.72	0.000061
Within Groups	1,864.45	39	47.81		
Total	3,791.53	45			

Annotations:

- (1) Sum of squared residuals from fitting the full (separate-means) model.
- (2) Sum of squared residuals from fitting the reduced (equal-means) model.
- (3) degrees of freedom.
- (4) Extra sum of squares is Total minus within (d.f. also obtained by subtraction).
- (5) A mean square is the ratio of a sum of squares to its degrees of freedom.
- (6) The *F*-statistic is the ratio of the Between MS to the Within MS.
- (7) The p-value comes from an *F*-distribution with 6 and 39 d.f.

Note: This is s_p^2 .

The table is organized so that all calculations follow a fixed pattern after entry of the residual sums of squares from the full (separate-means) and reduced (equal-means) models and their degrees of freedom. The between-groups sum of squares and its degrees of freedom are found by subtraction. Each mean square equals its sum of squares divided by its degrees of freedom. The *F*-statistic equals the ratio of the between-groups mean square to the within-groups mean square; and the evidence is summarized by the *p*-value, equal to the upper tail area from the *F*-distribution, whose numerator degrees of freedom are those associated with between groups and whose denominator degrees of freedom are those associated with within groups. The *p*-value of 0.000061 provides convincing evidence that at least one of the judges' means differs from one of the others' means.

5.4 MORE APPLICATIONS OF THE EXTRA-SUM-OF-SQUARES *F*-TEST

5.4.1 Example: Testing Equality in a Subset of Groups

The one-way analysis of variance *F*-test shown in Display 5.10 provided convincing evidence that the means were not equal for all seven judges. The next step in the analysis is to see whether the six other judges (not Spock's) have the same mean percentage of women on their venires. Although this step does not answer the main question of interest, it establishes an important context for addressing the main question. If there are no systematic differences among judges A through F in their inclusion of women, then the comparison of them to Spock's judge has greater relevance (in the court proceedings) than if they do differ among each other.

One way to test the hypothesis $H: \mu_2 = \mu_3 = \dots = \mu_7$ is to perform a one-way analysis of variance F -test with the venires from Spock's judge excluded from the analysis. This approach would be satisfactory for the needs of this problem. A better approach, however, includes Spock's judge's venires. It is "better" because it includes all available data in the pooled estimate of σ . Since this approach is also ideal for illustrating the general usefulness of the extra-sum-of-squares principle, it will be demonstrated here. It is based on a comparison of the following full and reduced models:

Group:	1	2	3	4	5	6	7
Full model (separate-means model):	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7
Reduced model (others-equal model):	μ_1	μ_0	μ_0	μ_0	μ_0	μ_0	μ_0

where μ_0 is used to represent the common mean of the last six judges in the reduced model. The estimated means are:

Group:	1	2	3	4	5	6	7
Full model (separate-means model):	\bar{Y}_1	\bar{Y}_2	\bar{Y}_3	\bar{Y}_4	\bar{Y}_5	\bar{Y}_6	\bar{Y}_7
Reduced model (others-equal model):	\bar{Y}_1	\bar{Y}_0	\bar{Y}_0	\bar{Y}_0	\bar{Y}_0	\bar{Y}_0	\bar{Y}_0

where \bar{Y}_0 is the average percentage among all venires for the other six judges.

The F -test for comparing these reduced and full models is not automatically computed in the analysis of variance program. Since the full model in this hypothesis is the separate-means model, its sum of squared residuals is available as the within-groups sum of squares in the analysis of variance table in Display 5.10 (1,864.45 on 39 degrees of freedom). The sum of squared residuals from the reduced model can be found by performing a second one-way classification analysis with just the two groups: *Spock* and *others*. The resulting residual sum of squares is 2,190.90, with $46 - 2 = 44$ degrees of freedom.

The extra-sum-of-squares F -statistic is $[(2,190.90 - 1,864.45)/(44 - 39)]/(1,864.45/39) = 1.37$, and the p -value is the proportion of an F -distribution on 5 and 39 degrees of freedom that exceeds 1.366, which is 0.26. There is consequently no evidence from these data of differences in means among the six other judges.

The next step in the analysis, assuming the six other judges do have equal means, is a test of the hypothesis that the Spock judge's mean is equal to the common mean of the other six. An F -test compares a full model with two means (one for the Spock judge and one for the other six) to a reduced model with a single mean:

Group:	1	2	3	4	5	6	7
Full model (others-equal model):	μ_1	μ_0	μ_0	μ_0	μ_0	μ_0	μ_0
Reduced model (equal-means model):	μ						

Notice that the previous step in the analysis has indicated the appropriateness of the two-parameter model, with a common mean for the other judges. This then becomes the full model for the inferential question: Is the Spock judge's mean different from the mean common to the six others, $H: \mu_1 = \mu_0$? The reduced model, consequently, is the equal-means model. The sum of squared residuals from

this reduced model is the total sum of squares from the one-way analysis of variance table: 3,791.53 on 45 degrees of freedom. The sum of squared residuals from this full model is 2,190.90 on 44 degrees of freedom. The F -statistic is $[(3,791.53 - 2,190.90)/(45 - 44)]/(2,190.90/44)$, which is equal to 32.14. By comparison to an F -distribution on 1 and 44 degrees of freedom, the p -value is found to be 0.000001. There is convincing evidence that the Spock judge's mean differs from the others.

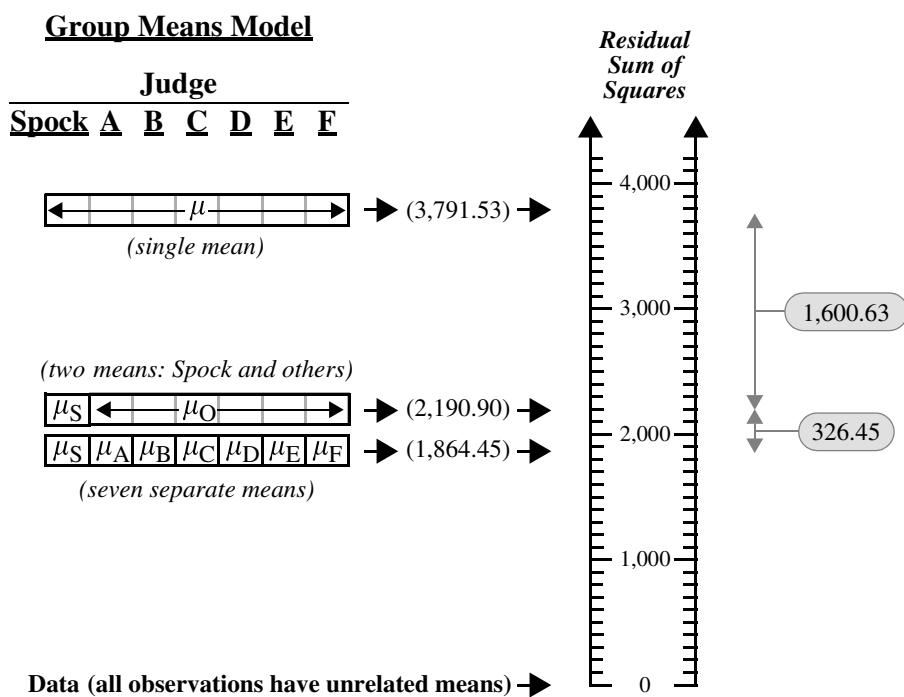
5.4.2 Summary of ANOVA Tests Involving More Than Two Models

The *others-equal* model (using two parameters to describe the means) lies intermediate between the *equal-means* model (with one parameter) and the *separate-means* model (with seven parameters). It is a simplification of the *separate-means* model, while the *equal-means* model is a simplification of it.

Display 5.11 illustrates the additive nature of the extra sums of squares for comparing these three nested models. The residual sum of squares in each model is viewed as a measure of the unexplained variability from that model, or as a *distance* from the model to the data. The residual sum of squares for the *equal-means* model is 3,791.53. By including one extra parameter to allow the Spock judge

DISPLAY 5.11

Residual sums of squares as distances from the data to proposed models for the means (Spock trial example)



DISPLAY 5.12

Complete analysis of variance table for three tests involving the mean percents of women in venires of seven judges

Source of variation	Sum of squares	d.f.	Mean square	F-statistic	p-value
Between groups	1,927.08	6	321.18	6.72	0.000061
Spock vs. others	1,600.63	1	1,600.63	33.48	0.000001
Among others	326.45	5	65.29	1.37	0.26
Within groups	1,864.45	39	47.81		
Total	3,791.53	45			

to have a different mean (the others-equal model), the unexplained variability is reduced by 1,600.63. By including an additional five extra parameters (the separate-means model), the unexplained variability is further reduced by 326.45. Notice that the extra sum of squares in the comparison of the separate-means model to the equal-means model (i.e., the between-group sum of squares) is the sum of the two-component sums of squares ($1,927.08 = 1,600.63 + 326.45$).

The two tests in the Spock trial data could be combined into a single analysis of variance table, as in Display 5.12. In this table, the *Between groups* sum of squares has been decomposed into two pieces corresponding to extra sums of squares for the two individual tests.

The *F*-statistic for “Spock vs. others” presented in Display 5.12 is *not* exactly the same as the extra-sum-of-squares *F*-statistic computed in Section 5.4.1, however. It has as its denominator the *Within groups* mean square, 47.81 (with 39 d.f.). The actual extra-sum-of-squares *F*-statistic has as its denominator the mean squared residuals from the *two-group* model, 49.79 ($=2,190.90/44$, with 44 d.f.). When presenting several tests simultaneously in a single analysis of variance table it is customary to use, as the denominator of the *F*-statistics, the estimate of σ^2 from the fullest model fit. There are some philosophical differences between these two approaches. For now, we encourage the student to apply the extra-sum-of-squares idea directly to address specific questions of interest (as in Section 5.4.1), but to be aware of differences in computer output due to the type of convention used in Display 5.12.

5.5 ROBUSTNESS AND MODEL CHECKING

5.5.1 Robustness to Assumptions

The robustness of *t*-tests, *F*-tests, and confidence intervals to departures from model assumptions can be described essentially in the same way as in Chapter 3:

1. Normality is not critical. Extremely long-tailed distributions or skewed distributions coupled with different sample sizes present the only serious distributional problems, particularly if sample sizes are small.

2. The assumptions of independence within and across groups are critical. If lacking, different analyses should be attempted.
3. The assumption of equal standard deviations in the populations is crucial.
4. The tools are not resistant to severely outlying observations.

The robustness with respect to equal standard deviations in the populations requires further discussion. It is important to pool together the estimates of variability from all the groups in order to make the most powerful comparisons possible. If one of the populations has a very different standard deviation, however, serious problems may result, even if the comparisons do not involve the mean from that population.

A computer was used to simulate situations in which three samples from normal populations were selected, with the aim of obtaining a 95% confidence interval for the difference in means from the first two. Six different configurations of population standard deviations were used with each of four sample size combinations. Based on 2,000 simulated data sets, the results appear in Display 5.13. The procedure is robust against unequal standard deviations for those situations where success rates are approximately equal to 95%.

DISPLAY 5.13

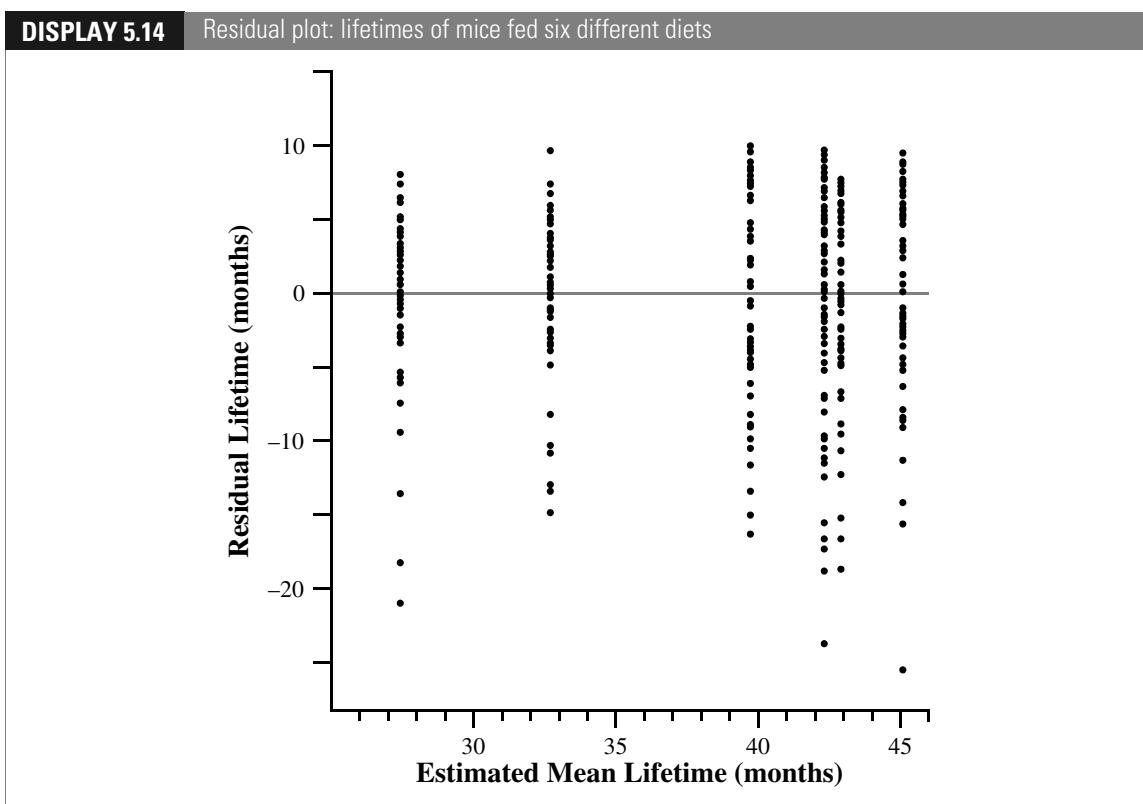
Success rates for 95% confidence intervals for $\mu_1 - \mu_2$ from samples simulated from normal populations with possibly different SDs

			$\sigma_2 = \sigma_1$			$\sigma_2 = 2\sigma_1$		
n_1	n_2	n_3	$\sigma_3 = \sigma_1$	$\sigma_3 = 2\sigma_1$	$\sigma_3 = 4\sigma_1$	$\sigma_3 = \sigma_1$	$\sigma_3 = 2\sigma_1$	$\sigma_3 = 4\sigma_1$
10	10	10	95.4	98.9	99.9	91.9	96.8	99.6
20	10	10	95.5	98.7	99.8	84.8	91.7	98.9
10	20	10	94.1	98.7	99.9	97.0	98.8	99.8
10	10	20	95.6	99.6	99.9	90.4	97.5	99.9

The simulations suggest that if σ_3 is quite different from σ_1 and σ_2 , then the actual success rates of 95% confidence intervals for $\mu_1 - \mu_2$ can be quite different from 95%. Unlike the result for the two-sample tools, the effect of unequal standard deviations can be serious even if the three sample sizes are equal.

5.5.2 Diagnostics Using Residuals

Initial graphical examination of the data by stem-and-leaf diagrams, box plots, or histograms is important. If there is a large number of groups, then side-by-side box plots are particularly useful, since they eliminate much of the clutter contained in other displays. As in the two-sample problem, initial assessment helps identify (1) the centers, (2) the relative spreads, (3) the general shapes of the distributions, and (4) the presence of outliers. If the spreads are quite different, transforming the data to a different scale should be considered.



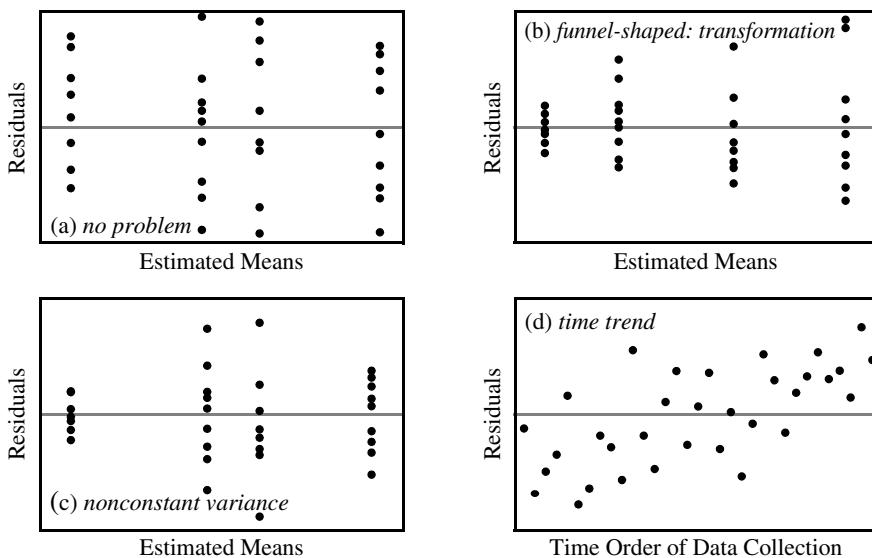
An important tool, which extends to almost all statistical methods in this book, is a *residual plot*. Because the residuals $\bar{Y}_{ij} - \bar{Y}_i$ are the original observations with their group averages subtracted out, they exhibit the variation of the observations without the visual interference caused by differences between group means. A scatterplot of these residuals versus the group averages can reveal a relationship between the spread and the group means, which may be exploited to improve the analysis.

Display 5.14 shows the residual plot from the diet restriction and longevity data. The features to look for in such a plot are (1) an increase in the spread from left to right, in a *funnel-shaped* pattern (which would suggest the need for a log or some other transformation), or (2) seriously outlying observations. The residual plot in Display 5.14 has neither feature, suggesting that analysis on the natural scale is adequate. It is evident that the distributions of lifetimes are skewed. For these large samples, however, there should be no problem in relying on the inferential tools derived from the normal model.

Finally, if the data were collected over time, a plot of residuals versus the time or order of data collection will help to reveal any serial effects that might be present. A pattern of generally increasing or generally decreasing residuals indicates a time trend, which may be accounted for using regression techniques. A pattern in which

DISPLAY 5.15

Some important patterns in residual plots



residuals close to each other tend to be more alike (or perhaps more different) than any two arbitrarily chosen residuals may indicate serial correlation. Formal investigation into serial correlation, and methods that account for it, are provided in Chapter 15.

Display 5.15 shows patterns in residual plots that would indicate (a) no obvious problems; (b) nonconstant variance, particularly a variance that is increasing with increasing mean (a log transformation might be in order); (c) nonconstant variance, particularly a variance that is smaller for small and large means (this might be the case if the response is restricted to be between 0 and 1, and a logit transformation might help: $\log[Y/(1 - Y)]$); and (d) a relationship between the response and time order of data collection.

5.6 RELATED ISSUES

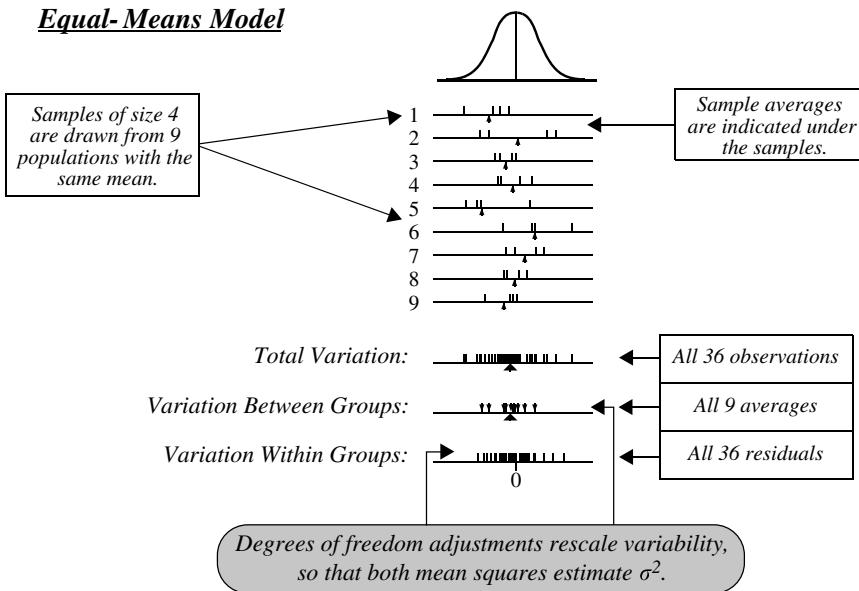
5.6.1 Further Illustration of the Different Sources of Variability

The analysis of variance is a general method of partitioning total variation into several components. This section attempts to shed further light on that partitioning.

A computer was used to generate independent random samples of size four each from nine normal populations having the same mean and the same standard deviation. The samples are illustrated in Display 5.16. The smooth curve above is the common population histogram, centered at the mean, μ . The horizontal axes

DISPLAY 5.16

Three sources of variation for data simulated from the equal-means model

Equal-Means Model

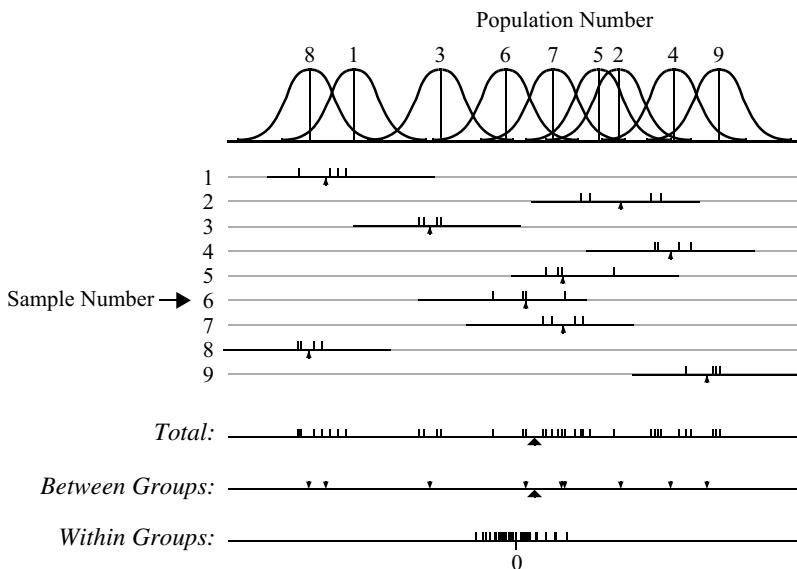
below the histogram locate the separate samples. The ticks above each axis locate the four sample values, and the single arrow below each axis locates the sample's average.

Three additional axes appear below the nine samples, labeled according to the three sources of variation in the analysis of variance table. The *Total* axis shows all 36 sample values on one common axis. Sample values are ticked on top of the axis and the grand average of all 36 is marked below the axis. On the *Between Groups* axis, bullets locate the nine sample averages on top, and their average—also the grand average—is marked by the arrowhead on the bottom. The *Within Groups* axis displays all 36 *residuals* from the separate-means model. These are the observations minus their group averages.

Variation When All Means Are Equal

Variation is different on the three different summary axes for two basic reasons: (1) *Observations are always closer to their sample average than to their population mean*. This is easily visible in sample #1, where all four observations fall below the population mean and where the sample average follows them. And (2) *sample averages are less variable than individual sample values*. Reason (2) explains why the between-group variation is visibly less than the total variation. Reason (1) explains why the within-group variation is also visibly less than the total variation.

The means of the sampling distributions of the average squared distance of a tick from the axis center are available from theory. The average of the squared residuals is $(1/36) \sum \sum (Y_{ij} - \bar{Y}_i)^2$. The mean of its sampling distribution is $(27/36)\sigma^2$.

DISPLAY 5.17 Variations in the several-group problem for data simulated from the separate-means model**Separate-Means Model**

This is less than σ^2 because of reason (1) mentioned above, and illustrates the need for a degrees of freedom adjustment: The mean square of the residuals, $[1/(36-9)] \sum \sum (Y_{ij} - \bar{Y}_i)^2$ is an unbiased estimate of σ^2 . (Its sampling distribution has mean σ^2 .)

Since the populations have the same mean, the nine \bar{Y}_i 's are like a sample of size 9 from a normal population with mean μ and variance $\sigma^2/4$. The quantity $\sum(\bar{Y}_i - \bar{Y})^2/9$ is an unbiased estimate of $(8/9)(\sigma^2/4)$. It is not surprising then that the sample variance of the sample of averages, $\sum(\bar{Y}_i - \bar{Y})^2/8$, is an unbiased estimate of $\sigma^2/4$. It follows that the between-group mean square, which is $\sum n_i (\bar{Y}_i - \bar{Y})^2/8$, with $n_i = 4$ for all i in this case, is an unbiased estimate of σ^2 .

If the hypothesis of equal population means is correct, the numerator and the denominator of the F -statistic (the between-groups and within-groups mean squares) are both unbiased estimates of σ^2 , so the F -statistic should be close to 1. As shown in the next display, if the population means are, in fact, unequal, then the numerator of the F -statistic is estimating something larger than σ^2 and the F -statistic will tend to be substantially larger than 1.

Variation When the Means Are Different

Display 5.17 depicts simulated samples drawn from populations with *different* means. The samples are the same as in Display 5.16, but shifted to their population means. Notice that the within-group variation is unchanged, whereas the

difference in population means results in a larger between-group variation. Hence the F -statistic will be larger than in the equal-means case.

Data that turn out like the bottom three axes in Display 5.17 show strong evidence that the between-group variability is much larger than that expected from the equal means model. The one-way analysis of variance F -test formalizes the comparison. In particular, while the denominator of the F -statistic is always an estimate of σ^2 , the numerator is estimating something larger than σ^2 . The amount by which it exceeds σ^2 depends on how different the means are. An F -statistic much larger than 1 provides strong evidence of differences among the population means.

5.6.2 Kruskal–Wallis Nonparametric Analysis of Variance

One method for coping with seriously outlying observations is to replace all observation values by their ranks in a single combined sample and then apply a one-way analysis of variance F -test on the rank-transformed data. The Kruskal–Wallis test, which is available in many statistical computer packages, is similar in its approach but takes advantage of the known variance of the ranks.

The Kruskal–Wallis test statistic is

$$KW = 1/[\sigma_R^2] \times \text{Between Group Sum of Squares (of ranks)},$$

where σ_R^2 is the variance of all n ranks (using an $n - 1$ divisor) and where n is the total number of observations in all groups. A p -value is found as the proportion of a chi-squared distribution on $(I - 1)$ degrees of freedom that is larger than this test statistic.

Display 5.18 shows the rank-transformed data from the Spock trial example. Notice that Spock's judge has venires that rank quite low. The between-group sum of squares (which could be obtained from an analysis of variance table based on the data in Display 5.18) is 3,956.325. The variance of the ranks is 180.122. The value of the Kruskal–Wallis test statistic is therefore 21.96, and p -value is 0.0012 from a chi-square distribution with 6 degrees of freedom. Testing equality of the

DISPLAY 5.18 Spock trial data, rank-transformed

Judge	Rank of venire from smallest (1) to largest (46) percentage of women								
Spock's	1	2	3	4	5	6	9.5	11	16
A	8	31	37	44	46				
B	22	26	34	36	41	45			
C	14	17	23.5	23.5	30	32.5	35	38.5	38.5
D	19	28							
E	9.5	12	15	25	40	43			
F	7	13	18	20	21	27	29	32.5	42

judges other than Spock's can be accomplished by performing a Kruskal–Wallis test ignoring the venires of the Spock judge. The Spock judge can be compared to the other judges combined using the rank-sum test.

5.6.3 Random Effects

Rationale for the Random Effects Model

It has so far been assumed that there is direct interest in the particular groups chosen. Sometimes, however, the groups are selected as representative of some broader population, and an inference is to be drawn to that population. A distinction is made between the *fixed effects model*, in which the group means are fixed and the *random effects model*, in which the group means are a random sample from a population of means. For the case studies in this chapter there is direct interest in the particular groups, so fixed effects models were used.

To illustrate when each model is appropriate, suppose that measurements are taken on the yield of a machine operated by each of several operators. An analysis of variance may be used to compare the mean yields under the different operators. A fixed effects model would be appropriate if there was interest in only those particular operators. (They may constitute all the operators at the plant.) A random effects model would be appropriate if those operators were just a sample and if the question of interest pertained to the population of operators from which they were sampled. There may be interest, for example, in estimating the proportion of the yield variance that could be explained by between-operator variability in a plant with a large number of operators.

Is the random effects model the right one to use? There are two pertinent questions: (1) Is inference desired to a larger set from which these groups are a sample? (2) Are the groups (operators) truly a random sample from the larger set? A yes answer to the first question would ordinarily prompt a user to use the random effects model. Statistical inference to the larger population would only be justified, however, if there was also a yes answer to the second question. If there was no random sampling to obtain the particular operators, then the usual warnings about potential biases due to using nonrandom samples apply.

The Random Effects Model

In the fixed effects model, observed sample i is thought to be a random sample from a normal population with mean μ_i and variance σ^2 . There are $I + 1$ parameters in the fixed effects model: the I means and the single variance σ^2 .

In the random effects model the μ_i 's themselves are thought to be a random sample from a normal population with mean μ and variance σ_μ^2 . The random effects model has three parameters: the overall mean μ , the within-group variance σ^2 , and the between-group variance σ_μ^2 .

Analysis of the one-way classification random effects model involves a test of whether σ_μ^2 is zero and an estimate of the ratio $\sigma_\mu^2 / (\sigma_\mu^2 + \sigma^2)$. Notice that this ratio is between 0 and 1. It is 0 when there is no between-group variance, and it is 1 when there is no within-group variance. Since the denominator describes the total

variance of the measurements, the ratio may be interpreted as the proportion of the total variance of the measurements that is explained by between-group variability. It is also called the *intraclass correlation*.

Estimation and Testing

The overall mean μ is estimated by the grand average \bar{Y} . The estimates of the two variances σ^2 and σ_μ^2 are often found by equating the mean squares in the analysis of variance table to the means of their sampling distributions, under the random effects model. In particular, letting $MS(W)$ and $MS(B)$ represent the within-group and between-group means squares, respectively, one obtains

$$\text{Mean}\{MS(W)\} = \sigma^2$$

$$\text{Mean}\{MS(B)\} = \sigma^2 + \frac{1}{n(I-1)} \left(n^2 - \sum_{i=1}^I n_i^2 \right) \sigma_\mu^2,$$

so the estimates of the variances are

$$\hat{\sigma}^2 = MS(W)$$

and

$$\hat{\sigma}_\mu^2 = \frac{n(I-1)[MS(B) - MS(W)]}{n^2 - \sum_{i=1}^I n_i^2},$$

with the modification that the latter is set to zero if the numerator turns out to be negative.

It is sometimes desired to test the hypothesis $H: \sigma_\mu^2 = 0$, against the alternative that it is greater than zero. It should be evident that this is analogous to the hypothesis that the means are all equal in the fixed effects model. The usual F -test for the fixed effects model (Section 5.3.2) is appropriate for this hypothesis as well.

The bottom line is that testing for between-group differences may be carried out in the usual way with an analysis of variance procedure. An additional, useful summary for the random effects model, however, is $\hat{\sigma}_\mu^2 / (\hat{\sigma}_\mu^2 + \hat{\sigma}^2)$.

Example—Spock Trial Data

Although the questions of interest in the Spock example focus on the specific judges, the data set can be used as an example to demonstrate random effects, by ignoring Spock's judge and thinking of the six other judges as representative of some large population. The parameters in the random effects model are μ , the overall mean percentage of women on venires, the variance σ^2 of percentages about the judge mean, and the variance σ_μ^2 of the population of judge means. A test of whether the between-judge variance is zero ($H: \sigma_\mu^2 = 0$) is the F -test from the standard analysis of variance. The p -value is 0.32, so the data are consistent with there being no between-judge variability. The estimates of σ_μ^2 and σ^2 are 1.96 and 53.6, so it can

be said that the proportion of variability that is due to differences between judges is $1.96/(1.96 + 53.6) = 0.035$. The square root of this number is also interpreted as the intraclass correlation—the correlation that percentages from two venires have if the venires come from the same judge. Of course, these inferential statements are speculative since the six judges were not, in fact, a random sample from a population of judges.

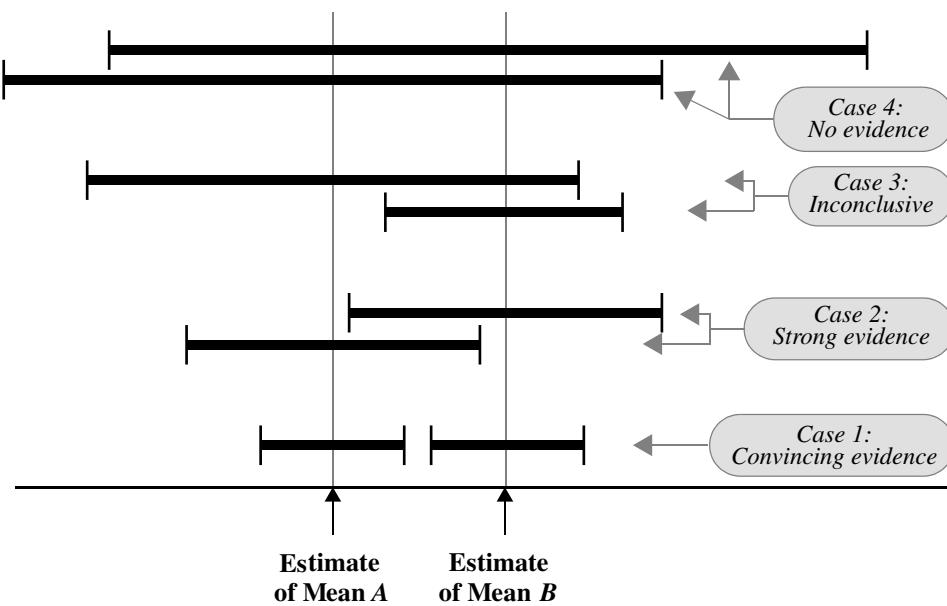
5.6.4 Separate Confidence Intervals and Significant Differences

Published research articles often present graphs of estimated means with separate confidence intervals for each. If a reader wishes to know whether two means are different, there is a fairly close—but not exact—relationship between the overlap in the confidence intervals and the result of a test of equal means. *The proper course of action for judging whether two means are equal is to carry out a t-test directly.* The comments here apply to a situation where either the reader is looking for a quick approximate answer or where the article fails to provide enough information to conduct the test.

Four categories of results are possible (Display 5.19). *Case 1:* If the intervals do not overlap, it is safe to infer that the means are different. Some readers incorrectly assume this is the only case that provides strong evidence of a difference. *Case 2,* however, also shows strong evidence. Even though the intervals overlap, the best estimate for each mean lies outside the confidence interval for the other mean.

DISPLAY 5.19

Separate confidence intervals for two group means: Are the means different?



Case 3, where one estimate lies within the confidence interval for the other mean, but the second estimate lies outside the first interval, is difficult to judge. But in *Case 4*, where the best estimate of each mean lies inside the confidence interval for the other, there is no evidence of any difference.

Finally, it must be mentioned that the discussion of this section applies to the comparison of two confidence intervals only. If there are more than two confidence intervals, then it may be quite misleading to compare the two most disparate ones, unless some adjustment for multiple comparisons is made. This topic is discussed in the next chapter.

5.7 SUMMARY

The term *analysis of variance* is often initially confusing as it seems to imply a comparison of variances. It is most definitely a method for comparing means, however, and the name derives from the approach for doing so—assessing variability from several sources. The analysis of variance *F*-test is used for assessing equality of several means.

Another point of confusion arises from the mistaken belief—due to the prevalence of the *F*-test in textbooks and computer programs—that the *F*-test necessarily plays a central role in the analysis of several samples. Usually it does not. It offers a convenient approach for detection of group differences, but it does not ordinarily provide answers to particular questions of interest. Tests and confidence intervals for pairs of means or linear combinations of means (discussed in the next chapter) provide much more specific information.

Analysis of data in several samples begins with a graphical display, like side-by-side box plots. Transformations of the data should be considered. The need for transformation and the presence of outliers is often better indicated by a residual plot—a plot of residuals versus fitted values. A funnel shape indicates the need for a transformation like the log, for example, for positive data. The analysis of variance table provides the numerical components of the *F*-test for equality of means. It also contains the within-groups mean square, which exactly equals the pooled estimate of variance, the best estimate of σ^2 . Confidence intervals and *t*-tests for pairs of means should use this pooled estimate of variance from all groups.

Diet Restriction and Longevity Study

In this study, the questions of interest called for five specific pairwise comparisons among the groups. It might be tempting to perform five two-sample *t*-tests, but it is a much more efficient use of the data to perform *t*-tests using a pooled estimate of variance from all the groups. The analysis begins with examination of side-by-side box plots. Although there is some skewness in the data, it is not enough to warrant concern—the tests are sufficiently robust against this type of departure from normality—and no transformation is suggested. A closer look at possible problems is available through a residual plot, but the spreads appear to be approximately equal and there are no serious outliers. The analysis proceeds, therefore, with *t*-tests

and confidence intervals in the usual way, but using the pooled estimate of variance from all groups.

Spock Trial Study

The stem-and-leaf plots in Display 5.4 and box plots in Display 5.5 are useful for suggesting some answers to the question of interest and for indicating the appropriateness of the tools based on the standard one-way classification model. An analysis of variance F -test confirms the strong evidence of some differences between means. An application of the extra-sums-of-squares F -test for comparing equality of the six other judges shows no evidence of a difference in mean percentages of women on their venires. Assuming that the six other means are equal, a further F -test shows overwhelming evidence that the Spock judge mean is different from the mean of the other six. Since this is a test for equality of two means, a t -test could be used. In fact, the F -test is equivalent to a two-sided t -test when $I = 2$. The actual p -value reported in the summary of statistical findings comes from a different test, not based on the assumption of equal means among the other six judges, and is discussed in the next chapter.

5.8 EXERCISES

Conceptual Exercises

1. **Spock Trial.** Why is it important to obtain a pooled estimate of variance in the Spock trial study? Is it ever a mistake to obtain a pooled estimate of variance in a comparison involving several groups?
2. Four methods of growing wheat are to be compared on five farms. Four plots are used on each farm and each method is applied to one of the plots. Five measurements are therefore obtained on yield per acre for each of the four growing methods. Are the methods of this chapter appropriate for analyzing the results?
3. **Diet Restriction.** Is there any explanation for why the distribution of lifetimes of mice in Display 5.1 are all negatively skewed?
4. **Diet Restriction.** For comparing group 3 to group 2, explain why it is better to use the t -tools presented in Section 5.2.3 (using s_p from all six groups) than to use the Chapter 2 t -tools (using s_p from only the two groups involved).
5. **Spock Trial.** Should Spock's accusers question the defense on how the venires were selected for their study?
6. **Spock Trial.** Why is it useful to test whether the six judges other than Spock's have equal mean percentages of women on their venires?
7. Why is s_p^2 not simply taken as the average of the I sample variances?
8. **Diet Restriction.** If the longevity study was a planned experiment, why are the sample sizes different?
9. If s_p is zero, what must be true about the residuals?
10. Explain the role of degrees of freedom of the F -distribution associated with the F -statistic. How are degrees of freedom related to how far the F -statistic is likely to be from 1?

11. What does it mean if the F -statistic is so *small* that the chance of getting an F -statistic that small or smaller is only, say, 0.0001?

12. Flycatcher Species Identification. One of the most challenging field identification problems for North American ornithologists is to distinguish the 10 species of flycatchers in the genus *Empidonax*. Many articles have appeared in popular and scientific journals suggesting different morphological clues to proper identification. F. Rowland ("Identifying *Empidonax* Flycatchers: The Ratio Approach," 2009, *Birding* 41 (2): 30–38) asserted that the relative size of wing length to tail length is the appropriate physical characteristic for distinguishing the species in the field. This conclusion was based on the average values of the wing length minus the tail length for 24 birds in each species, as shown in the following table.

Species:	Yellow-bellied	Acadian	Alder	Willow	Least	Hammond's	Gray	Dusky	Pacific-slope	Cordilleran
Average wing-tail (mm; $n=24$)	13.6	15.4	14.7	12.4	9.2	13.7	10.3	7.0	9.5	9.5

Explain why a conclusion that this measurement tends to differ in the 10 species cannot be made from the averages alone. What additional piece of information is needed to test for group differences and to evaluate the extent to which individuals from different species can be distinguished?

Computational Exercises

13. Spock Trial. By examining Display 5.8, answer the following:

- (a) What is the average percentage of women from all 46 venires?
- (b) For how many of the 9 Spock judge's venires is the percentage of women less than the grand average from all 46 venires?
- (c) For how many of the 9 Spock judge venires is the percentage of women less than the Spock judge's average?

14. Spock Trial. Use the following summary statistics to (a) compute the pooled estimate of the standard deviation and (b) carry out a t -test for the hypothesis that the Spock judge's mean is equal to the mean for judge A.

Judge:	Spock	A	B	C	D	E	F
Average % women:	14.62	34.12	33.61	29.10	27.00	26.97	26.80
SD of % women:	5.039	11.942	6.582	4.593	3.818	9.010	5.969
Sample size:	9	5	6	9	2	6	9

15. Spock Trial. (a) Use a calculator or statistical package to get the sample variance for the percentage of women on all 46 venires treated as one sample. (b) Multiply this by 45 to get the residual sum of squares for the equal-means model. (c) Multiply s_p^2 found in Exercise 14(a) above by $(46-7)$ to get the residual sum of squares for the separate-means model. (d) Use these to construct an analysis of variance table, including the F -statistic for the hypothesis of equal means. Compare the result with Display 5.10.

16. Spock Trial. Use a statistical computer package to obtain the analysis of variance table in Display 5.10.

17. Display 5.20 shows the start of an analysis of variance table. Fill in the whole table from what is given here. How many groups were there? Is there evidence that the group means are different?

DISPLAY 5.20 Incomplete ANOVA table for Exercise 17

Source	d.f.	Sum of squares	Mean square	F-statistic	p-value
Between groups	?	?	?	?	?
Within groups	24	35,088	?		
Total	31	70,907			

18. Fatty Acid. The data in Display 5.21 were obtained from a randomized experiment to estimate the effect of a certain fatty acid (CPFA) on the level of a certain protein in rat livers. Only one level of the CPFA could be investigated in a day's work, so a control group (no CPFA) was investigated each day as well. (Data from Donald A. Pierce.)

DISPLAY 5.21 Levels of protein ($\times 10$) found in rat livers

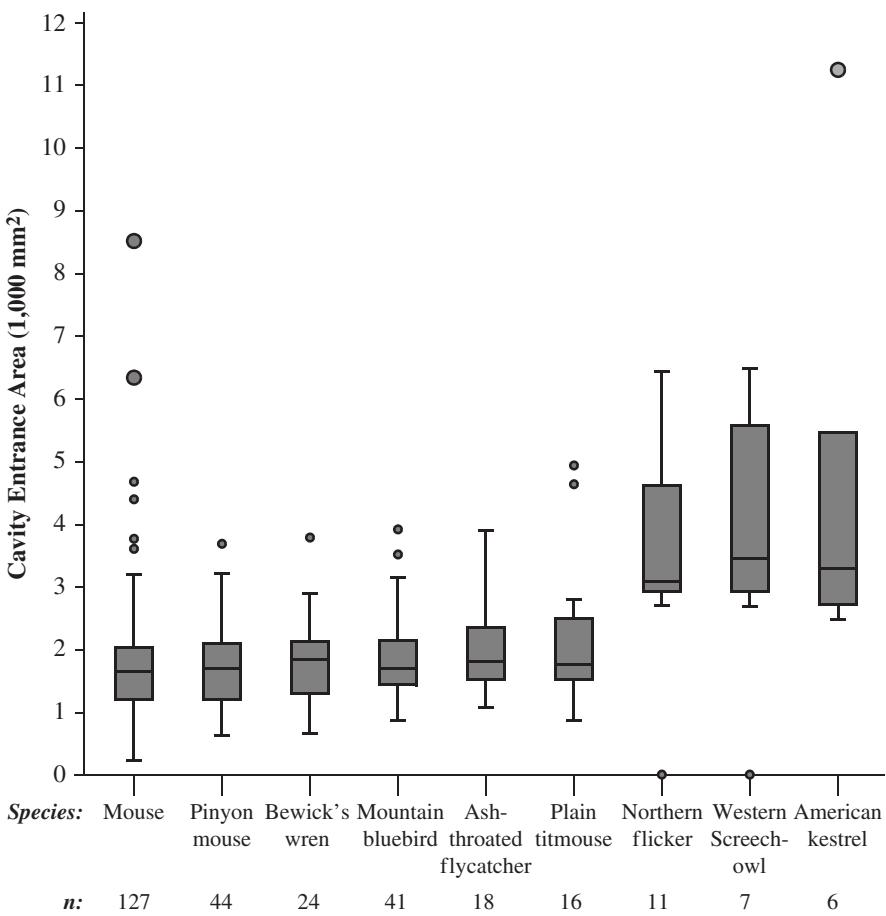
Day	Treatment					
	CPFA 50	CPFA 150	CPFA 300	CPFA 450	CPFA 600	Control
1	154, 177, 174					157, 165, 150
2		164, 192, 159				186, 206, 195
3			157, 159, 124			192, 202, 216
4				160, 152, 141		190, 187, 160
5					147, 152, 158	191, 188, 199

- (a) Obtain estimated means for the model with six independent samples, one for each treatment. Determine the residuals and plot them versus the estimated means. Plot the residuals versus the day on which the investigation was conducted. Is there any indication that the methods of this chapter are not appropriate?
- (b) Obtain estimated means for the model with 10 independent samples, one from each treatment-day combination. Calculate the ANOVA F -test to see whether these 10 groups have equal means.
- (c) Use (a) and (b) and the methods of Section 5.4.1, to test whether the means for the control groups on different days are different. That is, compare the model with 10 different means to the model in which there are 6 different means.

19. Cavity Size and Use. Biologists freely discuss the concept of competition between species, but it is difficult to measure. In one study of competition for nesting cavities in Southeast Colorado, Donald Youkey (Oregon State University Dept. of Fisheries & Wildlife) located nearly 300 cavities occupied by a variety of bird and rodent species. Display 5.22 shows box plots of the entrance area measurements from cavities chosen by nine common nesting species. The general characteristics—positive skewness, larger spreads in the groups with larger means—suggest the need for a transformation. On the logarithmic scale, the spreads are relatively uniform, and the summary statistics appear in Display 5.23. Are the species competing for the same size cavities? Or, are there differences in the cavity sizes selected by animals of different species? It appears that there are two very different sets of species here. The first six selected relatively small cavities while the last three selected larger ones. Is that the only significant difference?

DISPLAY 5.22

Box plots for areas of entrances to cavities used by different species



- (a) Compute the pooled estimate of variance.
- (b) Construct an analysis of variance table to test for species differences. (The sample standard deviation of all 294 observations as one group is $SD = 0.4962$.) Perform the F -test.
- (c) Verify that the analysis of variance method for calculating the between-group sum of squares yields the same answer as the formula

$$\text{Between-group SS} = \sum_{i=1}^I n_i \bar{Y}_i^2 - n \bar{Y}^2.$$

- (d) Fit an intermediate model in which the first six species have one common mean and the last three species have another common mean. Construct an analysis of variance table with F -tests to compare this model with (i) the equal-means model and (ii) the separate-means model. Perform the F -test.

DISPLAY 5.23

Summary statistics for areas of cavity entrances (logarithmic scale)

Species	<i>n</i>	Mean	Sample SD
Mouse	127	7.347	0.4979
Pinyon mouse	44	7.368	0.4235
Bewick's wren	24	7.418	0.3955
Mountain bluebird	41	7.487	0.3183
Ash-throated flycatcher	18	7.563	0.3111
Plain titmouse	16	7.568	0.4649
Northern flicker	11	8.214	0.2963
Western Screech-owl	7	8.272	0.3242
American kestrel	6	8.297	0.5842

20. Flycatcher Species Identification. Consider the table of averages (of wing lengths minus tail lengths) from 24 birds in each of 10 species of flycatcher in Exercise 12. If it is assumed that the 11 populations all have the same mean and same population standard deviation, what is an estimate of the population standard deviation?

21. A robust test for equality of several population variances is *Levene's test*, which was previously discussed in Section 4.5.3 for the case of two variances. This procedure carries out the usual one-way analysis of variance *F*-test on the absolute values of the differences of observations from their group medians. For practice, carry out Levene's test on the Spock data.

22. Equity in Group Learning. Several studies have demonstrated that engaging students in small learning groups increases student performances on subsequent tests. However, N. M. Webb and her colleagues argue that this raises a question of equity: Does the quality of the learning depend on the composition of the group? They chose students from five 7th and 8th grade classes in the Los Angeles school system. Based upon a science and language pretest, they classified each student's ability level as Low, Low-Medium, Medium-High, or High. They formed study groups consisting of three students each. The students were given a problem involving the setting up of two electrical circuits that would produce different brightness in a standard lightbulb. Each group was given the equipment to work with and time to discuss the problem and come to a solution. Afterward, each student was tested on the basics of the problem and its solution.

The table in Display 5.24 shows the results of the scores on this final test of the students whose ability level was Low in pretest. The students are grouped in the table according to the highest level of ability of a member in their study group. (Data from N. M. Webb, K. M. Nemer, A. W. Chizik, and G. Sugrue "Equity Issues in Collaborative Group Assessment: Group Composition and Performance," *American Educational Research Journal* 35(4): (1998) 607–51.)

DISPLAY 5.24

Achievement test scores of Low ability students who worked in different study groups

Highest ability level in the study group				
	Low	Low-medium	Medium-high	High
<i>Average:</i>	0.26	0.37	0.36	0.47
<i>St. Dev.:</i>	0.14	0.21	0.17	0.21
<i>n:</i>	17	24	25	14

DISPLAY 5.25

Achievement test scores of High ability students who worked in different study groups

	Lowest ability level in the study group			
	Low	Low-medium	Medium-high	High
Average:	0.75	0.77	0.72	0.85
St. Dev.:	0.16	0.11	0.12	0.10
n:	13	22	42	28

- (a) How strong is the evidence that at least one group mean differs from the others?
- (b) Display 5.25 shows a companion table. How strong is the evidence from this table that at least one mean differs from the others?
- (c) The study groups apparently were not formed using random assignment. How might this affect any conclusions you might draw from the analysis?

Data Problems

23. Was *Tyrannosaurus Rex* Warm-Blooded? Display 5.26 shows several measurements of the oxygen isotopic composition of bone phosphate in each of 12 bone specimens from a single *Tyrannosaurus rex* skeleton. It is known that the oxygen isotopic composition of vertebrate bone phosphate is related to the body temperature at which the bone forms. Differences in means at different bone sites would indicate nonconstant temperatures throughout the body. Minor temperature differences would be expected in warm-blooded animals. Is there evidence that the means are different for the different bones? (Data from R. E. Barrick, and W. J. Showers, "Thermophysiology of *Tyrannosaurus rex*; Evidence from Oxygen Isotopes," *Science* 265 (1994): 222–24.)

24. IQ and Future Income. Display 5.27 shows the first five rows of a data set with annual incomes in 2005 for 2,584 Americans who were selected in the National Longitudinal Study of Youth 1979, who were available for re-interview in 2006, and who had paying jobs in 2005, along with the quartile of their AFQT (IQ) test score taken in 1981 (see Exercise 2.22). How strong is the evidence that the

DISPLAY 5.26Measurements of oxygen isotopic composition of vertebrate bone phosphate (per mil deviations from SMOW) in 12 bones of a single *Tyrannosaurus rex* specimen

Bone	Oxygen isotopic composition				
Rib 16	11.10	11.22	11.29	11.49	
Gastralia	11.32	11.40	11.71		
Gastralia	11.60	11.78	12.05		
Dorsal vertebra	10.61	10.88	11.12	11.24	11.43
Dorsal vertebra	10.92	11.20	11.30	11.62	11.70
Femur	11.70	11.79	11.91	12.15	
Tibia	11.33	11.41	11.62	12.15	12.30
Metatarsal	11.32	11.65	11.96	12.15	
Phalange	11.54	11.89	12.04		
Proximal caudal	10.93	11.01	11.08	11.12	11.28
Mid-caudal	11.35	11.43	11.50	11.57	11.92
Distal caudal	11.95	12.01	12.25	12.30	12.39

DISPLAY 5.27

Annual income in 2005 and test score quartile for an IQ test taken in 1981 for 2,584 Americans in the NLSY79 survey; first 5 of 2,584 rows

Subject	IQquartile	Income2005
2	1stQuartile	5,500
6	4thQuartile	65,000
7	2ndQuartile	19,000
8	2ndQuartile	36,000
9	3rdQuartile	65,000

distributions of 2005 annual incomes differ in the four populations? By how many dollars or by what percent does the distribution of 2005 incomes for those within the highest (fourth) quartile of IQ test scores exceed the distribution for the lowest (first) quartile?

25. Education and Future Income. The data file ex0525 contains annual incomes in 2005 of a random sample of 2,584 Americans who were selected for the National Longitudinal Survey of Youth in 1979 and who had paying jobs in 2005 (see Exercise 22 in Chapter 2). The data set also includes a code for the number of years of education that each individual had completed by 2006: <12, 12, 13–15, 16, and >16. How strong is the evidence that at least one of the five population distributions (corresponding to the different years of education) is different from the others? By how many dollars or by what percent does the mean or median for each of the last four categories exceed that of the next lowest category?

Answers to Conceptual Exercises

1. To make comparisons, one must estimate variation. There are not many venires for any particular judge, so pooling the information gives better precision to the variance estimate. But if the groups have very different spreads, pooling is a bad idea.
2. Not appropriate. You should not expect the measurements from plots on the same farm to be independent of each other.
3. Perhaps there is something like an upper bound, a maximum possible lifetime for each group, and healthy mice all tend to get close to it. Unhealthy mice, however, die off sooner and at very different ages.
4. If the variances in all populations are equal, s_p from all groups uses much more data to estimate σ , resulting in a more precise estimator.
5. Yes. Perhaps these are just as good as random samples of all venires for each judge. If there was any bias in the selection, however—for example, if the nine venires for Spock's judge were chosen because they did not have many women—the results would be misleading.
6. Spock's lawyers will have a stronger case if they can show that Spock's judge is particularly different from *all others* in having low representation of women.
7. It is, if the sample sizes are all equal. Otherwise, it gives more weight to estimates from larger samples.
8. It is unusual for experimenters to purposefully plan on unequal sample sizes. In this study it is likely that the larger number of mice in the N/R50 group was planned, because that was the major experimental group. Inequalities in the other group sample sizes are likely the result of losing mice to factors unrelated to the experiment.

9. All the residuals would have to be identically zero for this to happen.
10. The larger the degrees of freedom in either the numerator or denominator, the less variability there is in their sampling distributions. With smaller degrees of freedom in either, sampling variability can result in an F -ratio which is considerably different from 1, even when the null hypothesis is true.
11. That would suggest that the sample averages are closer to each other than one would expect in the course of natural sampling from identical populations. You may want to check out the independence assumption.
12. There are two important quantities: (1) the within-mean square is s_p^2 , and (2) the p -value allows for judging group differences.

Linear Combinations and Multiple Comparisons of Means

The *F*-test for equality of several means gives a reliable result with any number of groups. Its weakness is that it neither tells which means are different from which others nor accounts for any structure possessed by the groups. Consequently, its role is mainly to act as an initial screening device.

If the groups have a structure, or if the research requires a specific question of interest involving several groups, a particular *linear combination* of the means may address the question of interest. This chapter shows how to make inferences about linear combinations of means and how to choose linear combinations for some important kinds of problems.

When no planned comparison is called for by the questions of interest or the group structure, one may compare all means with each other. The large number of comparisons, however, compounds the statistical uncertainty in the statements of evidence. Some methods of adjustment to account for this *multiple comparisons* problem are provided and discussed here.

6.1 CASE STUDIES

6.1.1 Discrimination Against the Handicapped—A Randomized Experiment

The U.S. Vocational Rehabilitation Act of 1973 prohibited discrimination against people with physical disabilities. The act defined a handicapped person as any individual who had a physical or mental impairment that limits the person's major life activities. Approximately 44 million U.S. citizens fit that definition. In 1984, 9 million were in the labor force, and these individuals had an unemployment rate of 7%, compared to 4.5% in the nonimpaired labor force.

One study explored how physical handicaps affect people's perception of employment qualifications. (Data from S. J. Cesare, R. J. Tannenbaum, and A. Dalessio, "Interviewers' Decisions Related to Applicant Handicap Type and Rater Empathy," *Human Performance* 3(3) (1990): 157–71.) The researchers prepared five videotaped job interviews, using the same two male actors for each. A set script was designed to reflect an interview with an applicant of average qualifications. The tapes differed only in that the applicant appeared with a different handicap. In one, he appeared in a wheelchair; in a second, he appeared on crutches; in another, his hearing was impaired; in a fourth, he appeared to have one leg amputated; and in the final tape, he appeared to have no handicap.

Seventy undergraduate students from a U.S. university were randomly assigned to view the tapes, fourteen to each tape. After viewing the tape, each subject rated the qualifications of the applicant on a 0- to 10-point applicant qualification scale. Display 6.1 shows the results. The question is, do subjects systematically evaluate qualifications differently according to the candidate's handicap? If so, which handicaps produce the different evaluations?

DISPLAY 6.1

Stem-and-leaf diagrams of applicant qualification scores given to applicants simulating five different handicap conditions

	None	Amputee	Crutches	Hearing	Wheelchair
0					
1	9	9		4	7
2	5	56		149	8
3	06	268	7	479	5
4	129	06	033	237	78
5	149	3589	18	589	03
6	17	1	0234	5	1124
7	48	2	445		246
8			5		
9					

Legend: 7 | 4 represents a score of 7.4 on the Applicant Qualification Scale.

Statistical Conclusion

The evidence that subjects rate qualifications differently according to handicap status is moderately strong, but not convincing (F -test p -value = 0.030). The difference between the average qualification scores given to the *crutches* candidate and to the *hearing-impaired* candidate is difficult to attribute to chance. The difference is estimated to be 1.87 points higher for the *crutches* tape, with a 95% confidence interval from 0.14 to 3.60 points based on the Tukey-Kramer procedure. The strongest evidence supports a difference between the average scores given to the *wheelchair* and *crutches* handicaps and the average scores given to the *amputee* and *hearing* handicaps (t -statistic = 3.19 for a linear contrast). None of the average qualification scores from the various feigned handicaps differ significantly from the no-handicap control. (The protected least significant differences all have two-sided p -values > 0.05 .)

Scope of Inference

Although the evidence suggests that differences exist among some of the handicap categories, the overall picture is made difficult by the location of the control in the middle of the groups. Any inference statements must also be qualified by the fact that the subjects used in this study may not accurately represent the population of employers making hiring decisions.

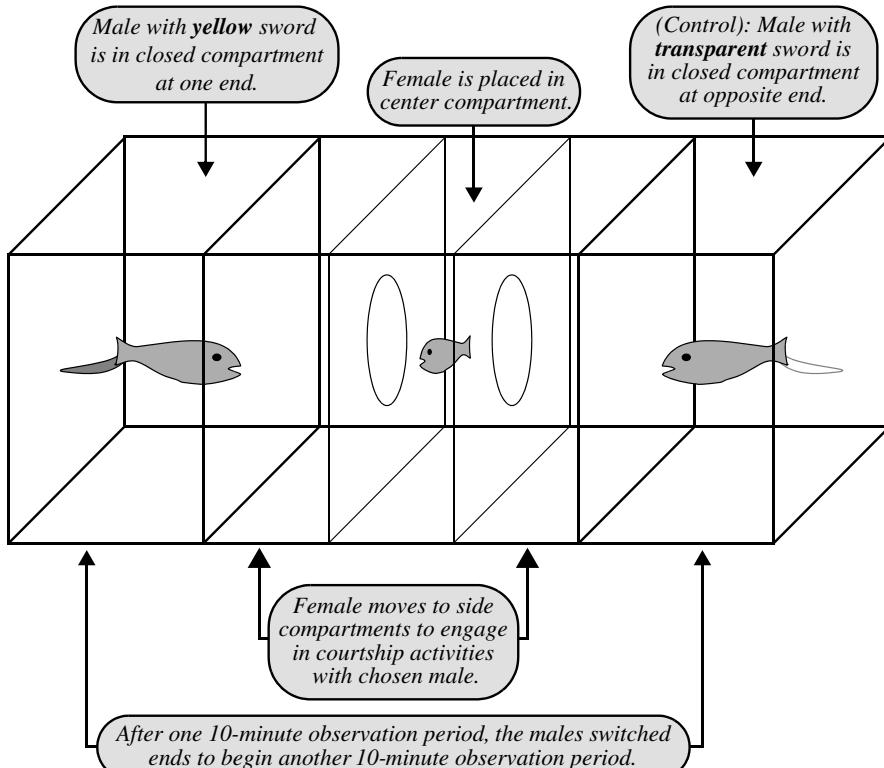
6.1.2 Pre-Existing Preferences of Fish—A Randomized Experiment

Charles Darwin proposed that sexual selection by females could explain the evolution of elaborate characteristics in males that appear to decrease their capacity to survive. In contrast to the usual model stressing the co-evolution of the female preference with the preferred male trait, A. L. Basolo proposed and tested a selection model in which females have a pre-existing bias for a male trait, even before males of the same species possess it. She studied a Central American genus of small fish. The males in some species of the genus develop brightly colored swordtails at sexual maturity. For her study, Basolo selected one species in the genus—the Southern Platypfish—whose males do not naturally develop the swordtails.

Six pairs of males were surgically given artificial, plastic swordtails. One male of each pair received a bright yellow sword, while the other received a transparent sword. The males in a pair were placed in closed compartments at opposite ends of a fish tank. One at a time, females were placed in a central compartment, where they could choose to engage in courtship activity with either of the males by entering a side compartment adjacent to it (see Display 6.2). Of the total time spent by each female engaged in courtship during a 20-minute observation period, the percentages of time spent with the yellow-sword male were recorded. These appear in Display 6.3. (Data from A. L. Basolo, “Female Preference Predates the Evolution of the Sword in Swordtail Fish,” *Science* 250 (1990): 808–10.) Did these females show a preference for the males that were given yellow swordtails?

DISPLAY 6.2

Experimental tank allowing female fish to choose between males

**Statistical Conclusion**

These data provide convincing evidence that the females tended to spend a higher percentage of time with the yellow-sword male than with the transparent-sword male (one-sided p -value < 0.0001 from a one-sample test that the mean percentage of time spent with the yellow-sword male is 50%). The estimated mean percentage of time with the yellow-sword male was 62% (95% confidence interval: 59% to 65%). The data provide no evidence that the mean percentage differed among the six pairs (p -value = 0.56, from a one-way analysis of variance F -test). There was also no evidence of a linear association between mean percentage of time spent with the yellow-sword male and the males' body size (p -value = 0.32 from a linear contrast).

6.2 INFERENCES ABOUT LINEAR COMBINATIONS OF GROUP MEANS

6.2.1 Linear Combinations of Group Means

Questions of interest sometimes involve comparing only two group means. Each question in the diet and lifetime study (Section 5.1.1) had this feature. Examining

DISPLAY 6.3

Percentage of courtship time spent by 84 females with the yellow-sword male; body sizes of the males are shown in parentheses

	Pair 1 (35 mm)	Pair 2 (31 mm)	Pair 3 (33 mm)	Pair 4 (34 mm)	Pair 5 (28 mm)	Pair 6 (34 mm)
	43.7	52.5	91.0	72.2	78.3	33.4
	54.0	65.6	62.0	58.5	66.0	42.2
	49.8	68.5	10.0	51.0	47.7	35.6
	65.5	45.9	83.8	56.8	77.5	79.9
	53.1	80.2	91.3	92.4	58.3	59.0
	53.0	67.0	56.3	55.3	61.1	58.1
	62.3	73.0	83.6	59.3	65.1	64.2
	49.4	71.7	53.3	42.0	62.9	82.8
	45.7	55.0	36.5	68.5	61.0	75.7
	56.6	70.0	65.4	78.4		66.3
	59.0	63.2	48.1	69.6		56.3
	67.8	39.6	50.6	89.2		84.5
	73.3	41.0	40.4	67.3		61.1
	43.8	59.2	90.6	77.5		87.6
	67.4		74.9			
	58.1		56.0			
			67.5			
Average:	56.41	60.89	62.43	67.00	64.21	63.34
SD:	9.02	12.48	22.29	14.33	9.41	17.68
n:	16	14	17	14	9	14

differences between the corresponding sample averages answers such questions. But this situation is uncommon in complex studies.

More typically, questions of interest involve several group means. For example, the study of qualification scores given to handicapped applicants might focus on a comparison of two handicaps—*crutches* and *wheelchair*—with the two handicaps—*hearing* and *amputee*. If $\mu_1, \mu_2, \mu_3, \mu_4$, and μ_5 are the mean scores in the *none*, *amputee*, *crutches*, *hearing*, and *wheelchair* groups, respectively, that question can be explored by studying the difference between the average of mean responses, $\gamma = (\mu_3 + \mu_5)/2 - (\mu_2 + \mu_4)/2$.

The parameter γ introduced here is called a *linear combination* of the group means. Linear combinations have the form

$$\gamma = C_1\mu_1 + C_2\mu_2 + \cdots + C_I\mu_I.$$

in which the coefficients C_1, C_2, \dots, C_I are chosen by the researcher to measure specific features of interest. In the handicap example, $C_1 = 0, C_2 = C_4 = -1/2$, and $C_3 = C_5 = +1/2$. These particular coefficients add to zero, which gives this linear combination the special designation of being a *contrast*.

6.2.2 Inferences About Linear Combinations of Group Means

The Estimate of a Linear Combination and Its Sampling Distribution

The same linear combination of sample averages, called g , is the natural estimate of the parameter γ . It is

$$g = C_1 \bar{Y}_1 + C_2 \bar{Y}_2 + \cdots + C_I \bar{Y}_I.$$

The sampling distribution of this estimate has mean γ . The standard deviation in the sampling distribution is given by the formula

$$\text{SD}(g) = \sigma \sqrt{\frac{C_1^2}{n_1} + \frac{C_2^2}{n_2} + \cdots + \frac{C_I^2}{n_I}},$$

which depends on the nuisance parameter σ . This assumes that the equal-spread model applies. The shape of the sampling distribution is normal if the individual populations are normal, and it is approximately normal more generally.

Standard Errors for Estimated Linear Combinations

The standard error for g is obtained by substituting the pooled estimate for σ in the formula for the standard deviation of g :

$$\text{SE}(g) = s_p \sqrt{\frac{C_1^2}{n_1} + \frac{C_2^2}{n_2} + \cdots + \frac{C_I^2}{n_I}}.$$

Two-sample comparisons are a special case. To compare the mean score of the crutches ratings with the mean score of the no handicaps ratings, for example, the investigator chooses $C_3 = +1$ and $C_1 = -1$, with $C_2 = C_4 = C_5 = 0$. The expression under the square root in the standard error reduces to the familiar sum of the reciprocals of the two group sample sizes. The SD is estimated from information in all groups, even when only two groups are being compared.

Inferences Based on the t-Distributions

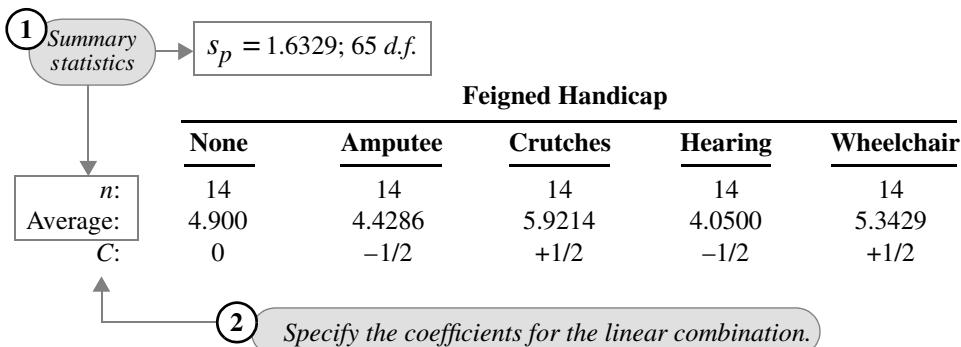
The t -ratio, $t = (g - \gamma)/\text{SE}(g)$, has an approximate Student's t -distribution with degrees of freedom equal to that of the pooled SD: d.f. = $(n_1 + n_2 + \cdots + n_I - I)$. The t -ratio may now be used as before either to construct a confidence interval or to test a hypothesized value for γ .

Example—Handicap Study

A computer can produce the averages and the pooled estimate of variability, but hand calculations are usually required from there. Display 6.4 illustrates all the steps

DISPLAY 6.4

Confidence interval construction for the linear combination $\gamma = (\mu_3 + \mu_5)/2 - (\mu_2 + \mu_4)/2$ in the handicap study



3 *Estimate the linear combination.*

$$g = \frac{(\bar{Y}_3 + \bar{Y}_5)}{2} - \frac{(\bar{Y}_2 + \bar{Y}_4)}{2} = \frac{(5.9214 + 5.3429)}{2} - \frac{(4.4286 + 4.0500)}{2} \\ = 1.3929$$

4 *Find the standard error of the estimate.*

$$\text{SE}(g) = 1.6329 \sqrt{\frac{(0)^2}{14} + \frac{(-1/2)^2}{14} + \frac{(1/2)^2}{14} + \frac{(-1/2)^2}{14} + \frac{(1/2)^2}{14}} \\ = 0.4364$$

5 *Construct the 95% confidence interval.*

$$t_{65}(0.975) = 1.9971 \quad \leftarrow \quad \text{from the t-distribution with } 65 \text{ d.f.}$$

$$1.3929 \pm (1.9971) \times (0.4364) \longrightarrow \text{from 0.521 to 2.264}$$

involved in finding a confidence interval for the contrast between the *wheelchair* and *crutches* means and the average of the *amputee* and *hearing* means. A 95% confidence interval for the parameter $\gamma = (\mu_3 + \mu_5)/2 - (\mu_2 + \mu_4)/2$ extends from 0.522 to 2.264.

6.2.3 Specific Linear Combinations

Here are some examples of linear combinations that arise frequently in practical situations.

Comparing Averages of Group Means

One common problem has two *sets of groups* distinguished by a specific factor; a comparison of the two sets is of interest. The preceding comparison is a typical example. Another example, from Section 5.1.2, involves comparing the Spock judge's mean percentage women with the average of the means from the other six judges.

If J groups in one set are to be compared to K groups in the second set, the relevant parameter is

$$\gamma = \frac{(\mu_{1,1} + \cdots + \mu_{1,J})}{J} - \frac{(\mu_{2,1} + \cdots + \mu_{2,K})}{K}.$$

The coefficients will be $+1/J$, $-1/K$, or zero, depending on whether a group is in the first set, the second set, or neither. *Important note:* One group cannot belong to both sets.

Comparing Rates

In problems like the diet restriction study in Section 5.1.1, where groups are structured according to levels of a quantitative explanatory variable, it may be desirable to report results as *rates* of increase in the mean response associated with changes in the explanatory variable. A comparison of the increase in mean lifetime associated with the reduction from 50 to 40 kcal/wk to the increase in mean lifetime associated with the reduction from 85 to 50 kcal/wk, for example, is best made on the basis of increases associated with one-unit changes in the caloric intake. Thus one would inquire whether the rate of increase in lifetime is the same in the reduction from 50 to 40 kcal/wk as in the reduction from 85 to 50 kcal/wk.

The rate of increase in mean lifetime associated with the reduction from 50 to 40 kcal/wk is $\text{rate2} = (\mu_6 - \mu_3)/(50 - 40)$, which is estimated to be

$$\begin{aligned}\text{est. rate2} &= \frac{(\text{Average lifetime on } N/R40) - (\text{Average lifetime on } N/R50)}{(50 - 40)} \\ &= \frac{(45.1 - 42.3)}{10} = 0.2800 \text{ months/[kcal/wk].}\end{aligned}$$

Similarly, the rate of increase associated with the reduction from 85 to 50 kcal/wk is $\text{rate1} = (\mu_3 - \mu_2)/(85 - 50)$, which is estimated to be:

$$\begin{aligned}\text{est. rate1} &= \frac{(\text{Average lifetime on } N/R50) - (\text{Average lifetime on } N/N85)}{(85 - 50)} \\ &= \frac{(42.3 - 32.7)}{35} = 0.2743 \text{ months/[kcal/wk].}\end{aligned}$$

Reducing caloric intake increased longevity in both instances, and the rates of increase appear to be about the same—about 0.28 month of extra lifetime for each kcal/wk reduction in caloric intake.

A formal comparison of the two rates will resolve whether the difference is real. The difference between the rates is estimated to be:

$$(\text{est. rate1} - \text{est. rate2}) = 0.2743 - 0.2800 = -0.0057 \text{ mo/[kcal/wk].}$$

This is a linear combination of only three means, because μ_3 occurs in both rates. To get the correct coefficients for calculating the standard error, reduce the comparison as follows:

$$\begin{aligned}\gamma &= (\text{rate1} - \text{rate2}) = \frac{(\mu_3 - \mu_2)}{35} - \frac{(\mu_6 - \mu_3)}{10} \\ &= -\frac{1}{35}\mu_2 + \frac{9}{70}\mu_3 - \frac{1}{10}\mu_6.\end{aligned}$$

This is a linear combination of three averages, so the standard error may be computed with the general formula of Section 6.2.2:

$$\begin{aligned}\text{SE}(g) &= (6.68) \sqrt{\frac{\left[-\frac{1}{35}\right]^2}{57} + \frac{\left[+\frac{9}{70}\right]^2}{71} + \frac{\left[-\frac{1}{10}\right]^2}{60}} \\ &= 0.1359 \text{ mo/[kcal/wk].}\end{aligned}$$

The estimate of σ , 6.68, is the pooled estimate from all six groups and has 343 degrees of freedom. The resulting t -statistic, $0.0057/0.1359 = 0.04$, provides no evidence that the two rates differ (two-sided p -value = 0.97). It might therefore be appropriate to estimate a common rate for increased lifetime over the entire 85 to 40 kcal/wk range.

Linear Trends

Sometimes the group means are associated with quantitative levels of an additional, *explanatory* variable. In the platyfish preference study, for example, the six groups correspond to different body sizes of the male pairs. A particular linear combination of group means may be used to assess the evidence for a linear trend in means as a function of the explanatory variable (for example, a linear trend in time with the yellow-sword male as a function of male body size).

Let X_i be the value of the explanatory variable associated with group i . Then the particular linear combination for linear trend happens to be the one that has $C_i = (X_i - \bar{X})$ as the coefficient of μ_i . The inference based on the linear combination will be unchanged if all the C_i 's are multiplied by the same constant. It is tidier, therefore, to use as C_i some multiple of $(X_i - \bar{X})$, where the multiplier is chosen to make all coefficients integers or to express the linear combination in convenient units of measurement. Display 6.5 demonstrates the linear combination for comparing the model of linear trend (in percentage time with the yellow-sword male), against the more general model in which the means are unrestricted. In this case a convenient multiplier to make all coefficients integers is 2, so that C_i is $2(X_i - \bar{X})$.

DISPLAY 6.5

Analysis of the pre-existing preference example: F -test for differences in mean percentage of time spent with yellow-sword male, and t -test for linear effect of male body size

ANOVA F-Test

Source of variation	Sum of squares	d.f.	Mean square	F-statistic	p-value
Between male groups	938.75	5	187.75	0.786	0.56
Within groups	18,636.68	78	238.93		
Total	19,575.43	83			

Conclusion: There is no evidence that the group means are different for different pairs of males (p -value = 0.56, from ANOVA F -statistic).

 t -Test for linear effect of body size

Group	n	Average (%)	Standard deviation	Male body size (mm)	Coefficient
Pair 1	16	56.41	9.02	35	5
Pair 2	14	60.89	12.48	31	-3
Pair 3	17	62.43	22.29	33	1
Pair 4	14	67.00	14.33	34	3
Pair 5	9	64.21	9.41	28	-9
Pair 6	14	63.34	17.68	34	3
Pooled	84	62.13	15.46	Average = 32.5	

① Calculate the coefficients for the linear combination.

$$C_i = 2*(X_i - 32.5)$$

② Calculate the effect's estimate

$$\begin{aligned} g &= (5)(56.41) + (-3)(60.89) + (1)(62.43) + (3)(67.00) + (-9)(64.21) + (3)(63.34) \\ &= -25.06 \end{aligned}$$

and its standard error.

$$\begin{aligned} SE(g) &= (15.46) \sqrt{\frac{(5)^2}{16} + \frac{(-3)^2}{14} + \frac{(1)^2}{17} + \frac{(3)^2}{14} + \frac{(-9)^2}{9} + \frac{(3)^2}{14}} \\ &= 54.77 \end{aligned}$$

③ Calculate the t -statistic and determine the p -value.

$$t\text{-statistic} = \frac{-25.06}{54.77} = -0.458 \quad \text{one-sided } p\text{-value} = 0.32 \quad (\text{from } t\text{-distribution with 78 d.f.})$$

Conclusion: There is no evidence of a linear association between group means and male body size (one-sided p -value = 0.32).

There are two tests shown in Display 6.5. The analysis of variance F -test indicates the evidence against the reduced, single-mean model as a special case of the full, six-mean model. The t -test for linear effect indicates the evidence against the reduced model in which the slope of a straight line function of body size is zero, as a special case of the model in which the slope is unrestricted.

Important note: It is useful to understand the linear combination for testing linear effect. For practical purposes, however, *regression methods* introduced in Chapter 7 provide a more complete data analytic process for investigating this type of structure.

Averages

Given the absence of any apparent differences between the mean percentages of times spent with the various yellow-sword males, the question of whether female platyfish prefer yellow-sword males becomes sensible globally. The average of the group means is another linear combination (but not a contrast), and the appropriate null hypothesis is that $(\mu_1 + \mu_2 + \dots + \mu_6)/6 = 0.5$. The average of the group sample averages is 62.38%. Its standard error, 1.72%, is calculated in the same way from the general formula. This leads to the conclusive statement that the females were not dividing their time evenly between the males.

Caveat concerning average: The overall average just described differs slightly from the grand average of all female percentages (treated as a single group), because it gives equal weight to each male pair instead of equal weight to each female percentage. The conclusions will, in most cases, agree.

6.3 SIMULTANEOUS INFERENCES

A 95% confidence interval procedure is successful in capturing its parameter in 95% of its applications. When several 95% confidence intervals are considered simultaneously, they constitute a *family* of confidence intervals. The relative frequency with which all of the intervals in a family simultaneously capture their parameters is smaller than 95%. Because this rate is often of interest, the following distinction is drawn.

Individual confidence level is the success rate of a procedure for constructing a single confidence interval.

Familywise confidence level is the success rate of a procedure for constructing a family of confidence intervals, where a “successful” usage is one in which all intervals in the family capture their parameters.

If the family consists of k confidence intervals, each with individual confidence level 95%, the familywise confidence level can be no larger than 95% and no smaller

than $100(1 - 0.05k)\%$. The actual familywise confidence level depends on the degree of dependence between the intervals.

The lower limit decreases rapidly with k . With a family of 10 confidence intervals, the familywise level could be as low as 50%, indicating the strong possibility that at least one of the intervals fails to capture its parameter. In other words, one should not suppose that all the confidence intervals capture their parameters, especially when a large number of intervals are being considered simultaneously.

The issue here is *compound uncertainty*—the increased chance of making at least one mistake when drawing more than one direct inference. Compound uncertainty also arises when many *tests* are considered simultaneously. The greater the number of tests performed, the higher the chance that a low p -value will be found for at least one of them, even in the absence of group differences. Consequently, the researcher is likely to find group differences that are not really there.

Multiple Comparisons

Multiple comparison procedures have been developed as ways of constructing individual confidence intervals so that the familywise confidence level is controlled (at 95%, for example). The important issue for the researcher to consider is whether to control the individual confidence levels or the overall confidence level.

Planned Comparisons, Unplanned Comparisons, and Data Snooping

Consider a one-way classification with 100 groups. One researcher may be particularly interested in comparing groups 23 and 78 because the comparison answers a research question directly. The researcher knows which groups are involved before seeing the data, and the comparison will be reported regardless of its statistical and practical significance. This constitutes a *planned comparison*. The individual confidence level should be controlled for planned comparisons.

Another researcher may examine differences between all possible pairs of groups—4,950 confidence intervals in all. As a result of these efforts, the researcher finds that groups 36 and 44 and groups 27 and 90 suggest actual group differences. Only these pairs are reported as significant. They exemplify *unplanned comparisons*. The familywise confidence level should be controlled for unplanned comparisons, since the uncertainty measure must incorporate the process of searching for important comparisons.

A third researcher notices that group 36 has the largest average and group 44 has the smallest, and presents only the single confidence interval—the one comparing group 36 to group 44. This is an instance of *data snooping*, in which the particular hypothesis or comparison chosen originates from looking at the data. The familywise confidence level should be controlled, for the same reason as in the second case.

The Spock trial, the diet restriction, and the platyfish preference studies all involved planned comparisons. The handicap study had no prespecified comparisons, so any comparisons reported in it should be treated as unplanned.

6.4 SOME MULTIPLE COMPARISON PROCEDURES

Confidence intervals for differences between pairs of means are centered at the difference between sample averages. Interval half-widths are computed as follows:

$$\text{Interval half-width} = (\text{Multiplier}) \times (\text{Standard error}).$$

As usual the standard error of the difference is the pooled standard deviation times the square root of the sum of reciprocals of sample sizes.

There are many multiple comparison procedures, and these differ in their multipliers. The two highlighted in the ensuing subsections offer strict control over the familywise confidence levels for two important families.

6.4.1 Tukey–Kramer Procedure and the Studentized Range Distributions

The Tukey–Kramer procedure utilizes the unique structure of the multiple comparisons problem by selecting a multiplier from the *studentized range distributions* rather than from the *t*-distributions. The idea is to incorporate the search for the two most divergent sample averages directly into the statistical procedure.

Consider the case where all group means are equal and where all sample sizes are equal. The standard errors for all comparisons are the same and are equal to SE, say. A confidence interval is successful when it includes zero, so success occurs for a particular comparison when the magnitude of the difference between sample averages, $|\bar{Y}_i - \bar{Y}_j|$, is small. All such differences are less than $M \times \text{SE}$ if $(\bar{Y}_{\max} - \bar{Y}_{\min})$, the range of sample averages, is less than $M \times \text{SE}$. By selecting M in such a way that the chance of getting

$$(\bar{Y}_{\max} - \bar{Y}_{\min}) \leq M \times \text{SE}$$

is 95%, one guarantees that all intervals include zero 95% of the time. That is, the overall familywise confidence level is set at 95%.

A studentized range distribution describes values for the ratio of the range in I sample averages to the standard error of a single sample average, given that all samples are drawn from the same normal population. Tables of the studentized range distributions provide the $100 \times (1 - \alpha)$ th percentile, $q_{I,\text{d.f.}}(1 - \alpha)$, depending on the number of groups (I) and the degrees of freedom (d.f.) for estimating σ . (In the one-way classification problem, d.f. = $n - I$.) The procedure originally proposed by Tukey—called Tukey’s HSD, for “honest significant difference”—assumed an ideal normal model with equal spreads and also assumed equal sample sizes in all groups. The modification for unequal sample sizes provides confidence intervals with approximately the correct confidence levels, and goes by the name of the Tukey–Kramer procedure. The multiplier used in the interval half-width calculation is $[q_{I,n-1}(1 - \alpha)]/\sqrt{2}$.

In the handicap study, $I = 5$ groups and $(n - I) = 65$ degrees of freedom. The 95th percentile in the corresponding studentized range distribution (interpolated

between 60 and 120 d.f. in Table A.5) is 3.975, so the multiplier for constructing 95% confidence intervals is 2.8107.

6.4.2 Dunnett's Procedure

Dunnett proposed a multiple comparison procedure for comparing every other group to a reference group. This is often appropriate, such as for randomized experiments that compare several other treatments to a control. Dunnett realized that the t -statistics for comparing $I - 1$ other groups to a reference group were correlated due to the common appearance of the reference group average. To achieve a familywise error rate, the Dunnett procedure replaces the usual t -distribution with a multivariate t -distribution, which accounts for that correlation. Since there are fewer comparisons in this family than in the family of all possible pairwise comparisons, the Dunnett-adjusted confidence intervals of a given confidence level will be narrower than those from the Tukey–Kramer procedure.

Tables for the multivariate t -distribution aren't readily available, but are incorporated in computer routines for multiple comparisons with the Dunnett procedure. Such a routine was used to find the multiplier for 95% confidence intervals for comparing every other group to “None” (i.e., control) in Case Study 6.1.1. to be 2.5032. Notice that this falls between the unadjusted t -multiplier, 1.9971, and the Tukey–Kramer multiplier, 2.8107.

6.4.3 Scheffé's Procedure

Scheffé proposed the multiplier

$$\sqrt{(I - 1)F_{(I - 1), \text{d.f.}}(1 - \alpha)},$$

where $F_{(I - 1), \text{d.f.}}(1 - \alpha)$ is the $(1 - \alpha)$ th percentile of the F -distribution with $I - 1$ and d.f. degrees of freedom. Here, $(I - 1)$ represents the between-group degrees of freedom, and d.f. is the within-group degrees of freedom. Scheffé's multiplier controls the overall confidence level for the family of parameters consisting of all possible linear contrasts among group means. When applied to the smaller family of differences between pairs of group means, the overall confidence level is *at least* $100(1 - \alpha)\%$, and generally is higher. The Scheffé method finds a more appropriate application in providing intervals for regression curves, as will be seen in later chapters.

In the handicap study, $I - 1 = 4$, d.f. = $n - I = 65$, and the 95th percentile in the F -distribution is 2.513. The resulting multiplier is 3.1705.

6.4.4 Other Multiple Comparisons Procedures

The multiple comparisons procedures in this section present a range of options for balancing individual and familywise confidence levels.

The LSD

The familiar choice for a multiplier is the $100(1 - \alpha/2)\%$ critical value in the Student's t -distribution with degrees of freedom equal to those associated with the pooled SD. The resulting interval half-width is called the *least significant difference*, or LSD. The terminology arises naturally because any difference that exceeds the LSD in size is significant in a $100\alpha\%$ -level hypothesis test. The multiplier for the handicap study is $t_{65}(0.975) = 1.9971$.

F-Protected Inferences

The method known as "protected LSD" is a simple and widely used alternative for testing unplanned comparisons. It is a two-step procedure, as follows:

1. Perform the ANOVA F -test.
2. (a) If the p -value from the F -test is large (>0.05 , say), do not declare any individual difference significant, even though some differences appear large enough to be declared real.
(b) If the p -value from the F -test is small, proceed with individual comparisons, as in Section 5.2, using t -tests or confidence intervals with the t -multiplier.

In the handicap study, the p -value from the ANOVA test was 0.03, so the F -protected comparison plan would proceed to step 2(b).

Although one should compute confidence intervals whether or not the F -test's p -value is small, there is no convenient method for F -protecting confidence intervals. It is tempting to use a t -multiplier in 2(b) and to substitute either the Tukey–Kramer or Scheffé multiplier in 2(a), depending on the family. That practice, however, would control neither the individual nor the familywise success rate at 95%.

Bonferroni

If the confidence level for each of k individual comparisons is adjusted upward to $100(1 - \alpha/k)\%$, the chance that all intervals succeed simultaneously is at least $100(1 - \alpha)\%$. This result is an application of the Bonferroni inequality in probability theory. Using the Student's t -multiplier $t_{d.f.}(1 - \alpha/2k)$ allows the user to be "at least $100(1 - \alpha)\%$ confident" that all intervals succeed. In a multiple comparisons problem involving I groups, there are $k = I(I - 1)/2$ pairs of means to be compared. The exact confidence level for the Bonferroni intervals is not generally as predictable as for the Tukey–Kramer intervals in the multiple comparisons problem. Bonferroni intervals may be used in a wider range of problems, however, including some situations where the Tukey–Kramer approach is not appropriate. With the five groups in the handicap study, $k = 10$, so the t -multiplier is the 99.75th percentile in the t -distribution with 65 d.f., or 2.9060.

Others

The *Newman–Keuls* procedure also employs studentized range distributions, but with different multipliers for different ranges. *Duncan's multiple range* procedure

DISPLAY 6.6

Summary of 95% confidence interval procedures for differences between treatment means in the handicap study

Group	Average	Difference with . . .			
		Hearing	Amputee	Control	Wheelchair
Crutches	5.921	1.871	1.492	1.021	0.578
Wheelchair	5.343	1.293	0.914	0.443	
Control	4.900	0.850	0.471		
Amputee	4.429	0.379			
Hearing	4.050				
Procedure		95% interval half-width			
LSD		1.233			
Dunnett		1.545 (for comparisons with control only)			
Tukey–Kramer		1.735			
Bonferroni		1.794			
Scheffé		1.957			

A confidence interval is centered at a difference with half-width given by one of the procedures.

extends the Newman–Keuls procedure, with a Bonferroni protective correction to the nominal level.

6.4.5 Multiple Comparisons in the Handicap Study

The pooled estimate of the standard deviation of the data in Display 6.1 is 1.633. All groups have the same sample size (14) so the standard error for any and all differences between sample averages is

$$\text{SE}(\bar{Y}_i - \bar{Y}_j) = 1.633 \sqrt{\frac{1}{14} + \frac{1}{14}} = 0.6172.$$

The other relevant information is as follows: There are $I = 5$ groups, so there are $k = 10$ different comparisons, and the degrees of freedom for the standard error are 65.

Display 6.6 summarizes the ten 95% confidence intervals computed according to the multiple comparisons methods described earlier. Since sample sizes are all the same, confidence intervals have the same width for all comparisons under each method. The upper part of Display 6.6 shows the centers of the confidence intervals. The lower part shows the interval half-widths.

If the researchers' intent is to compare each of the handicap groups to the control group, then the Dunnett procedure is appropriate. In this case, none of the 95% confidence intervals for the handicap minus control difference exclude zero, so the data provide no evidence of a handicap effect. If, however, their intent is to compare every group to every other group, then the Tukey–Kramer procedure is appropriate. It suggests one difference—between the *hearing* and *crutches* groups.

Recall, however, that comparing the combined *crutches* and *wheelchair* group with the combined *amputee* and *hearing* group (Display 6.4) revealed a very clear difference. Making that comparison was suggested by an examination of the data; by analogy, the interval here should also be widened by using the Scheffé multiplier in place of the *t*-multiplier. When this is done, the confidence interval for the contrast is from 0.011 to 2.775, which still excludes zero. Because the Scheffé method incorporates the search among linear contrasts for the most significant, one should conclude that strong evidence exists that this difference is real.

6.4.6 Choosing a Multiple Comparisons Procedure

The LSD is the most liberal procedure (narrowest confidence intervals), and the Scheffé procedure is the most conservative (widest confidence intervals). Bonferroni and Tukey–Kramer procedures occupy intermediate positions. Aside from conducting these general comparisons, the best approach is to think carefully about whether it is desirable to control the familywise confidence level and, if so, what the appropriate family of comparisons includes. If the answer to the first question is no, then standard *t*-tools apply. If the family of comparisons includes all differences of other groups with a reference group, the Dunnett method gives the appropriate control. If the family of comparisons includes pairwise differences of all group means, the Tukey–Kramer method gives precise control. If the family includes a large number of comparisons and no other method seems feasible, the Bonferroni method—although conservative—can always be applied.

6.5 RELATED ISSUES

6.5.1 Reasoning Fallacies Associated with Statistical Hypothesis Testing and *p*-Values

p-values indicate the *strength of evidence* from data in support of an alternative to a null hypothesis. Although *p*-value reasoning is both natural and logical, widespread misinterpretations and misuses cause many scientists and statisticians to advocate their elimination from scientific discourse. The abuses are easily understood, though. For the sake of preserving a tool that is very useful when used correctly, it is important for students of statistics to recognize and avoid the misinterpretations. To this end, Display 6.7 lists the important ones. Scientists that understand this display are unlikely to error in their interpretations of *p*-values.

6.5.2 Example of a Hypothesis Based on How the Data Turned Out

Although not a one-way classification problem, the following example demonstrates the need to incorporate data snooping into the assessment of uncertainty. The letters in Display 6.8 represent 2,436 mononucleotides in a DNA molecule. Mononucleotides come in four varieties—A, C, G, and T—and their sequence along the DNA strand forms a molecule's genetic code. DNA molecules break, drift for a time, and

DISPLAY 6.7

Common fallacies of reasoning from statistical hypothesis tests

Fallacy name	The fallacy	Avoiding the fallacy
False Causality Fallacy	Incorrectly interpreting statistical significance (i.e., a small p -value) from an observational study as evidence of causation	Use the word <i>association</i> to indicate a relationship that is not necessarily a causal one.
Fallacy of Accepting the Null	Incorrectly interpreting a lack of statistical evidence that a null hypothesis is false (i.e., a large p -value) as statistical evidence that the null hypothesis is true	Avoid this incorrect wording: “the study provides evidence that there is no difference.” Say instead: “there is no evidence from this study of a difference.” Also, report a confidence interval to emphasize the many possible hypothesized values (in addition to 0) that are consistent with the observed data.
Confusing Statistical for Practical Significance	Interpreting a “statistically significant” effect (which has to do with the strength of evidence that there’s an effect) as a practically important one, which it may or may not be	If you must use the term <i>statistically significant</i> , don’t abbreviate it. Also, report a confidence interval so that the <i>size</i> of an effect can be evaluated for its practical importance.
Data Dredging (Fishing for Significance, Data Snooping)	Incorrectly drawing conclusions from an unadjusted p -value that emerged from a process of sifting through many possible p -values Note: “Publication Bias” is the de facto data dredging that results if journals only accept research papers with statistically significant findings.	For multiple comparisons of means, use the adjustments in this chapter. For identifying a few from many possible predictor variables, use the variable selection methods in Chapter 12. For tests based on many different response variables (data mining), use the False Discovery Rate methods of Chapter 16.
Good Statistics from Bad Data	Incorrectly accepting conclusions based on sound statistical technique when there are problems with data collection, such as biased sampling or data contamination	Critically evaluate the potential biases from non-randomly selected samples.

DISPLAY 6.8

2,436 mononucleotides along a DNA molecule. All 40 occurrences of the trinucleotide TGG appear in boldface. Eleven breaks occurred in the string, at the positions indicated by dashes.

```

TAAAGAACATAATGCCGATATTGTTAATACTGTGTACTGTAAGAACATATTAGCATTGT
CTATGACTAAGAACATTCAAACATTATTGATGCTATAGGGTGGCAATATAATAGTCATTCT
TACGATATTGAAAAAGTTATCTCCTACTTCGACACATTACGTCAAAAATACACGAAA
ATAAAGATCCAGTTACTTGGGTTGCTAGACCTTGACATTACAGTTAACCTCTATAGTT
ATTTACGTATACTGGAAAGGTATATAATAGTCATAACCGTCATTAACTTATTTA-CGTGC
TTCTATATTAACTCTGAGAATTATCATCTACACTGTGATAAACTTATCTTACGAGATT
TAGAAAGGAATATTGTGTCGAGTGTACATGATTGG-TATATAATACGGACTATCCAATCTC
TTATGTCATACTAAAACCTTATTGCCAACACTTGGAACTTGGAAGATGACATCATAG
ACAATTGATTATCTATGAAACCTTATTCTAGAAAGGATGAAACTAAATGTTCCAGAT
GAGGATTATGAGTTGATTITGCTTAACTGTGGTATATAAACGCAAGAAATAATTAGGAA
ATCTGCTACTCCTATCAAAATGTAATCAGGTCAAATTATCTTCTCCCAGAGGTATAAAT
AATGTAACTGGATACTAGACTGTACCA-AAATATTCTTGTGATAAAACAACCACGCAA
TCATACAAGTATCCATTCACTAGATGATCTCATGAAACATTGGATCAAATTATAGATATATCCA
TATGTCATAAGTACTCATTGGGAGAAGTAGTATATCTCATCGGT-GATGGATGAAACAA
TGAAATACATAACAATGCTATAGCGTAATTATATCAAAACAATTTGGAT-TCCAATTCC
CCGATGAAATGCCAACGACTGTAGCTAGCGGGATACCCGCTAA-ATAAATTATACGTAG
TAGGAGGTCTACCAAAATCCACATCTGTTGAGCTGG-TCCACGGGGATGCTGCTTG
TTAATATGCCGAGTCTCTGAAACCTAGATGTAATCCAGCAGTGGC-ATCCATAAAACAATGT
TATATACTGTAATTGGAGGAGACATTCTGAAACTGATACAACAGAACATATTGCTACCCAA
CATGTCAGTGCAGTTGACCATTCCACTTATCTCATTATAAAATCATGCGCGTTAG
TGTTCGGTAGAAGGTTATTCTTGGTTGTTGAGAAATCGGAAATTATTGTGAATCCAGCAA
TACATGGCTCTGATAGATGATCCTATTATCCGAGGGATAATCCAGAATTGATCATAGTGG
ATAATAACTGCTATTGATAGGAGGATTAATCGTCATCGTATATAGATACTAGAAGT
GTACATCATCACACTTATTGAAATATTGGATTGGTAATAATTGAAATAAAAT
TAGTTTATGTCACATGAATTAAAC-TCACCGATTAGTTGTTAAGGAAACTAACAGA
GCTAAATCTCTACTAGGCAATCACCTACGCCCGGATATGATTATATAGCGCTTACGA
TTATCATCCCTCAGGAGAACGACAGTAAATTAAAGACAGATTAGTATGTCCATGCCT
AAGTTCTGCTATTGGTAGAATAGCTCTAGGTCTGTCTGCCTAAAGGCATTGATATAG
GAGGCGGTGTAATAGACGAAGATTATAGGGAAACATAGGAGTCATTCTATTAAATATG
GAAAATGTACGTTAATGTAATATGGAGGATACAATAGCTAGCTAACTATCAACGTAT
ATATTATCCAGAACTGGAGAACTGGACTACAATCTAGATATGGAAATAGGAGGAGATCAAGG
GTTT-GATCAACAGGACTTAGATAATAAACATAGTATGTTGTCGATGTTATAGTGTAA
AATATCGTAGATTATGATAGTATAGATAATTGGTATAGTACAGGATATAAGAGATGAG
GCTAGCAATAATGTTGATCACGACTATGTATATCCACTTCCAGAAAATTGGTATATAGAT
TTGACAAAGTCCACTAACATACTCGATTATCTATCACAGGAAACGGGACCATGTAATGGC
GTGTCGACTATATGAGTAACAAACAGTTAGACGACTTGTATAGACAGTTGCCAACAG
ACTAGATCATATAGATATTCAACATATATTGTGATAAAAGTTAGTAATGATTATAATAG
GGACATGAATATCATGTATGA-TATGGCATCTACAAACATTACAGTTATGACATAAAAT
AACGAAGTTAATACTATACTATGATAACAAAGGGGTTGGGTGAAGATTGCGACAATT
TCATTATAACCGAATTGGGTAGACGATGTATGA

```

then recombine with other strands to form new molecules. The molecule shown in Display 6.8 has broken in 11 places, indicated by the dashes between two mononucleotides. Three consecutive mononucleotides form a trinucleotide, which functions as a genetic word. In this molecule, the 40 occurrences of the trinucleotide TGG are shown in boldface. In line 7, a break point has appeared shortly after an occurrence of TGG, which is of interest because TGG may indicate an increased likelihood of a break to follow.

If a break occurs between any of the mononucleotides of TGG plus the four following, the break is said to be *downstream* from TGG. And in this particular molecule, 6 of the 11 breaks were downstream from a TGG trinucleotide. The question is, does this evidence support TGG as a precursor of breaks in the DNA?

Summary of Statistical Analysis

Two different analyses are possible, depending on whether the hypothesis that TGG was specifically indicated arose prior to viewing this molecule (planned comparison)

or whether TGG was actually suggested by examination of this molecule (data snooping).

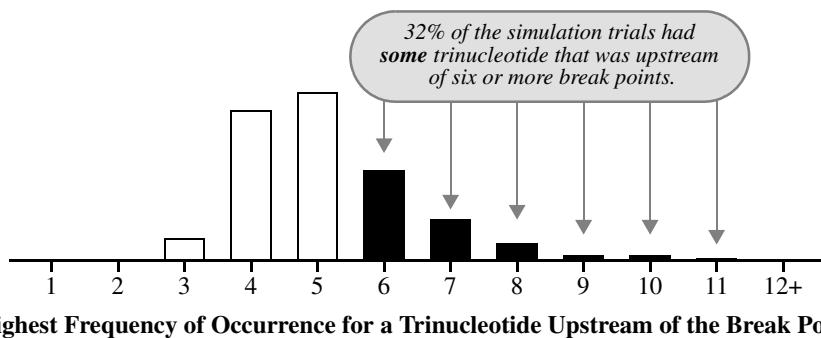
1. Of the 2,435 possible break positions in the entire string, 235 (9.65%) are downstream from TGG trinucleotides. If breaks occurred at *random* positions, about 9.65% of the 11 breaks (that is, one) would be downstream from TGG. But in fact, 6 of the 11 (54.55%) occurred downstream from TGG. That would seem to be too many to have happened by chance. If breaks occurred at random positions, the possibility that six or more would occur at positions downstream from TGG trinucleotides is precisely 0.000243. This is very small, showing strong evidence of an association between the occurrences of breaks and TGG.

The p -value of 0.000243 assumes, however, that, in every trial determining 11 breaks, the number downstream from TGG is counted. What if the focus on TGG were the result of a search on this molecule for the trinucleotide occurring most frequently upstream of the existing breaks? *If the search for the trinucleotide was an integral part of the statistical analysis of this data set, the series of trials used to evaluate the evidence must also include the search procedure.*

2. The computer was programmed to conduct a trial incorporating the search. It randomly selected 11 break points from the 2,435 possible, and searched upstream from all the breaks to find the trinucleotide appearing most frequently. Then it recorded the frequency of the occurrence of this trinucleotide. The computer ran 1,000 of these trials. Of course, the most frequently occurring trinucleotide differed considerably from trial to trial. The key piece of evidence is the distribution of the frequency of the most frequent upstream trinucleotide, which appears in Display 6.9. In 320 of the 1,000 trials, some trinucleotide was found upstream from the 11 breaks 6 times or more. In one trial, the same trinucleotide occurred upstream 11 times. Thus, having some trinucleotide upstream from many of the breaks in a molecule appears to be rather common! Consequently, if the search was conducted on this molecule, the p -value for an observed highest frequency of 6 is $p = 0.32$.

DISPLAY 6.9

Simulated estimate of the distribution of the highest frequency of occurrence of any trinucleotide upstream of 11 randomly selected breaks



Scope of Inference

This example illustrates the tremendous difference between evaluating the strength of evidence about a preplanned hypothesis and evaluating the strength of evidence about a hypothesis suggested by the data themselves. Correct evaluation of *post hoc* hypotheses is possible, but it must incorporate the actual procedure by which the hypothesis was formulated. (Data from M. B. Slabaugh, N. A. Roseman, and C. K. Mathews, “Amplification of the Ribonucleotide Reductase Small Subunit Gene: Analysis of Novel Joints and the Mechanism of Gene Duplication in Vaccinia Virus,” Biochemistry and Biophysics Department Report, Oregon State University, Corvallis, Oregon (1989).)

How valuable is the correct *p*-value? These researchers did “discover” TGG from this molecule. Had they accepted the strength of evidence 0.000243, they might have wasted considerable effort devising explanations for what turned out to be a nonrepeatable phenomenon.

6.5.3 Is Choosing a Transformation a Form of Data Snooping?

The Statistical Sleuth emphasizes examining graphical displays for clues to possible data transformation. When a stem-and-leaf diagram suggests a logarithmic transformation, should the resulting inferences take the data-selected transformation into account? The answer is yes. Uncertainty about which transformation (if any) is best is part of the problem, so a measure of uncertainty regarding the question of interest should incorporate uncertainty about the form of the statistical model.

In practice, however, most researchers do not incorporate model uncertainty into uncertainty measures—largely because doing so is extremely difficult. In later chapters (in particular, Chapter 12), *The Statistical Sleuth* will investigate some methods that have been devised for solving this problem. For now, the scope of inference should be limited to replicates where the transformation is selected. To caution a reader about additional uncertainty, the statistical summary should explain clearly all the steps taken to select the transformation.

6.6 SUMMARY

Many researchers routinely examine an analysis of variance *F*-statistic for group differences; if it is significant, they proceed directly to multiple comparison methods to search for differences. In most cases, this is a mistake. Usually the groups have an inherent structure related to levels of specific factors set by the researcher. Linear combinations of group means can be devised to answer questions related to this structure.

A host of studies fall within the general class of one-way classifications. Yet the statistical analysis is not the same for all. A major distinction must be made between studies in which the group structure calls for specific planned comparisons and those in which the only question of interest is which means differ from which

others. Selecting an appropriate statistical procedure requires the researcher to evaluate honestly whether hypotheses were clearly stated prior to data collection or whether the data themselves guided the formulation of hypotheses. In the latter case, proper statistical evaluation must acknowledge the data-snooping process.

This chapter demonstrates some statistical tools for assessing uncertainty when a family of inferences is desired. Selecting an appropriate tool requires some introspection about the nature of the family examined. This chapter also introduced a powerful tool—computer simulation—that can help evaluate evidence about more complex hypotheses suggested by the data.

6.7 EXERCISES

Conceptual Exercises

1. **Handicap Study.** (a) Is it possible that the applicant's handicap in the videotape is confounded with the actor's performance? (b) Is there a way to design the study to avoid this confounding?
2. **Mate Preference of Platfish.** If $\mu_1, \mu_2, \dots, \mu_6$ represent the mean percentage of time spent by females with the yellow-sword male, for the six pairs of males, (a) state the null and alternative hypotheses that are tested by (i) the analysis of variance F -test and (ii) the t -test for the hypothesis that the linear contrast (for the linear effect of male body size) is zero. (b) Say why it is possible that the second test might find evidence that the means are different even if the first does not.
3. **Mate Preference of Platfish.** For the test that the mean percent of time females spent with yellow-sword males is 50%, a one-tailed p -value was reported. Why?
4. An experimenter takes 20 samples of bark from each of 10 tree species in order to estimate the differences between fuel potentials. The data give 10 species averages, the lowest being 1.6 Btu/lb and the highest 3.8 Btu/lb for a range of 2.2 Btu/lb. A colleague suggests that another species be included, so the experimenter plans to gather 20 samples from that species and calculate its average potential. Which of the following is true about the range that the 11 species averages will have when the new species is included? (a) The range will equal the old range, 2.2 Btu/lb. (b) The range will be larger than 2.2 Btu/lb. (c) The range will be smaller than 2.2 Btu/lb. (d) The range cannot be smaller than 2.2 Btu/lb. (e) The range cannot be larger than 2.2 Btu/lb. (f) It is not possible to say that any of the above options is true until the average is known.
5. **O-Ring Data.** The case study in Section 4.1.1 involved the numbers of O-ring events on U.S. space shuttle flights launched at temperatures above and below 65°F. In the context of this chapter, is anything suspicious about that data? (*Hint:* Is there a possibility of data snooping?)
6. Does a confidence interval for the difference between two groups use information about variability from other groups? Why? or Why not?
7. What is the distinction between planned and unplanned comparisons?
8. Does a planned comparison always consist of estimating the difference between the means in two groups?
9. In comparing 10 groups a researcher notices that \bar{Y}_7 is the largest and \bar{Y}_3 is the smallest, and then tests the hypothesis that $\mu_7 - \mu_3 = 0$. Why should a multiple comparison procedure be used even though there is only one comparison being made?
10. If the analysis of variance screening test shows no significant evidence of any group differences, does that end the issue of there being any differences to report?

DISPLAY 6.10 Test scores for the experimental CAD instruction course

Group	Logo	Teaching method	n	Average	SD
1	L + D	Lecture and discussion	9	30.20	3.82
2	R	Programmed text	9	28.80	5.26
3	R + L	Programmed text with lectures	9	26.20	4.66
4	C	Computer instruction	9	31.10	4.91
5	C + L	Computer instruction with lectures	9	30.20	3.53

11. When choosing coefficients for a contrast, does the choice of $\{C_1, C_2, \dots, C_I\}$ give a different t -ratio than the choice of $\{3C_1, 3C_2, \dots, 3C_I\}$?

Computational Exercises

12. **Handicap Study.** Consider the groups *amputee*, *crutches*, and *wheelchair* to be handicaps of mobility and *hearing* to be a handicap affecting communication. Use the appropriate linear combination to test whether the average of the means for the mobility handicaps is equal to the mean of the communication handicap.
13. **Handicap Study.** Use the Bonferroni method to construct simultaneous confidence intervals for $\mu_2 - \mu_3$, $\mu_2 - \mu_5$, and $\mu_3 - \mu_5$ (to see whether there are differences in attitude toward the mobility type of handicaps).
14. **Handicap Study.** Examine these data with your available statistical computer package. See what multiple comparison procedures are available within the one-way analysis of variance procedure. Verify the 95% confidence interval half-widths in Display 6.6.
15. **Comparison of Five Teaching Methods.** An article reported the results of a planned experiment contrasting five different teaching methods. Forty-five students were randomly allocated, nine to each method. After completing the experimental course, a one-hour examination was administered. Display 6.10 summarizes the scores on a 10-minute retention test that was given 6 weeks later. (Data from S. W. Tsai and N. F. Pohl, "Computer-Assisted Instruction Augmented with Planned Teacher/Student Contacts," *Journal of Experimental Education*, 49(2) (Winter 1980–81): 120–26.)
- (a) Compute the pooled estimate of the standard deviation from these summary statistics.
 - (b) Determine a set of coefficients that will contrast the methods using programmed text as part of the method (groups 2 and 3) with those that do not use programmed text (1, 4, and 5).
 - (c) Estimate the contrast in (b) and compute a 95% confidence interval.
16. A study involving 36 subjects randomly assigned six each to six treatment groups gives an ANOVA F -test with p -value = 0.0850. What multipliers are used to construct 95% confidence intervals for treatment differences with the following methods: (i) LSD, (ii) F -protected LSD, (iii) Tukey–Kramer, (iv) Bonferroni, and (v) Scheffé?
17. **Adder Head Size.** Red Riding Hood: "My, what big teeth you have!" Big Bad Wolf: "The better to eat you with, my dear." Are predators morphologically adapted to the size of their prey? A. Forsman studied adders on the Swedish mainland and on groups of islands in the Baltic Sea to determine if there was any relationship between their relative head lengths (RHL) and the body size of their main prey, field voles. (Data from A. Forsman, "Adaptive Variation in Head Size in *Vipera berus* L. Populations," *Biological Journal of the Linnean Society* 43 (1991): 281–96.) Relative head length is head length adjusted for overall body length, determined separately for males and females. Field vole body size is a combined measure of several features, expressed on a standardized scale.

DISPLAY 6.11

Average relative head lengths of adders from seven μ Swedish localities with their distances to the mainland and the body sizes of prey

Locality	Sample size	Average relative head length	Distance (km) to mainland	Field vole body size
Uppsala	21	-6.98	0	-1.75
In-Fredeln	34	-4.24	25.1	
Inre Hamnskär	20	-2.79	13.4	-0.16
Norrpada	25	2.22	14.7	1.31
Kärringboskär	7	1.27	10.0	
Ängskär	82	1.88	22.7	1.67
Svenska Högarna	48	4.98	39.6	2.17

The data appear in Display 6.11. The pooled estimate of standard deviation of the RHL measurements was 11.72, based on 230 degrees of freedom.

- (a) Determine the half-widths of 95% confidence intervals for all 21 pairwise differences among means for the seven localities, using (i) the LSD method and (ii) the Tukey–Kramer method.
- (b) Using a linear contrast on the groups for which vole body size is available, test whether the locality means (of relative head length) are equal, against the alternative that they fall on a straight line function of vole body size, with nonzero slope.
- (c) Repeat (b) for the pattern of distances to the mainland rather than vole body size.

18. Nest Cavities. Using the nest cavity data in Exercise 5.19, estimate the difference between the average of the mean entry areas for flickers, screech-owls, and kestrels and the average of the mean entry areas for the other six animals (on the transformed scale). Use a contrast of means.

19. Diet Restriction. For the data in Display 5.1 (and the summary statistics in Display 5.2), obtain a 95% confidence interval for the difference $\mu_3 - \mu_2$ using the Tukey–Kramer procedure. How does this interval differ from the LSD interval? Why is the Tukey–Kramer procedure the wrong thing to use for this problem?

20. Equity in Group Learning. [Continuation of Exercise 5.22.] (a) To see if the performance of low-ability students increases steadily with the ability of the best student in the group, form a linear contrast with increasing weights: $-3 = \text{Low}$, $-1 = \text{Low-Medium}$, $+1 = \text{Medium-High}$, and $+3 = \text{High}$. Estimate the contrast and construct a 95% confidence interval. (b) For the High-ability students, use multiple comparisons to determine which group composition differences are associated with different levels of test performance.

21. Education and Future Income. Reconsider the data problem of Exercise 5.25 concerning the distributions of annual incomes in 2005 for Americans in each of five education categories. (a) Use the Tukey–Kramer procedure to compare every group to every other group. Which pairs of means differ and by how many dollars (or by what percent)? (Use p -values and confidence intervals in your answer.) (b) Use the Dunnett procedure to compare every other group to the group with 12 years of education. Which group means apparently differ from the mean for those with 12 years of education and by how many dollars (or by what percent)? (Use p -values and confidence intervals in your answer.)

- 22.** Reconsider the measurements of oxygen composition in 12 dinosaur bones from Exercise 5.23. Using a multiple comparisons procedure in a statistical computer package, find 95% confidence intervals for the difference in means for all pairs of bones (a) without adjusting for multiple comparisons and (b) using the Tukey–Kramer adjusted intervals. (c) How many of the unadjusted intervals exclude zero? (d) How many of the Tukey–Kramer adjusted intervals exclude zero? (e) By how much does the width of the adjusted interval exceed the width of the unadjusted interval, as a percentage, for comparing bone 1 to bone 2?

Data Problems

- 23. Diet Wars.** To reduce weight, what diet should one combine with exercise? Most studies of human dieting practices have faced problems with high dropout rates and questionable adherence to diet protocol. Some studies comparing low-fat and low-carbohydrate diets have found that low-carb diets produced weight loss early, but that the loss faded after a short time. In an attempt to exert more control on subject adherence, a team of researchers at Ben-Gurion University in Negev, Israel, conducted a trial within a single workplace where lunch—the main meal of the day—was provided by the employer under the guidance of the research team. The team recruited 322 overweight employees and randomly assigned them to three treatment groups: a low-fat diet, a low-carb diet (similar to the Atkins diet), and a Mediterranean diet. Trans-fats were discouraged in all three diets. Otherwise, the restrictions and recommendations were as follows:

	Low-fat (n = 104)	Mediterranean (n = 109)	Low-carbohydrate (n = 109)
Calorie/day restriction	Women: 1,500 kcal Men: 1,800 kcal	Women: 1,500 kcal Men: 1,800 kcal	(Not Specified)
Percentage of calories from fat	30%	35%	(not specified)
Carbohydrates/day	(not specified)	(not specified)	20 g at start; increasing to 120 g
Percentage of calories from saturated fat	10%	(not specified)	(not specified)
Cholesterol/day	300 mg	(not specified)	(not specified)
Recommended:	low-fat grains vegetables fruits legumes	30–45 g olive oil 5–7 nuts < 20 g vegetables fish and poultry	get fat and protein from vegetables
Discouraged:	added fats	beef and lamb sweets high-fat snacks	

The study ran for two years, with 272 employees completing the entire protocol. Display 6.12 shows some of a data set simulated to match the weight losses (kg) of the participants at the study's conclusion. Is there evidence of differences in average weight loss after two years among these diets? If so, which diets appear to be better than which others? (Notice the consequences of controlling the family-wise confidence level on the widths of 95% confidence intervals.)

- 24. A Biological Basis for Homosexuality.** Is there a physiological basis for sexual preference? Following up on research suggesting that certain cell clusters in the brain govern sexual behavior, Simon LeVay (data from S. LeVay, "A Difference in Hypothalamic Structure Between Heterosexual and Homosexual Men," *Science*, 253 (August 30, 1991): 1034–37) measured the volumes of four cell groups in the interstitial nuclei of the anterior hypothalamus in postmortem tissue from 41 subjects at autopsy from seven metropolitan hospitals in New York and California. The volumes of one cell

DISPLAY 6.12

Partial listing of data from a diet and weight loss experiment, showing subject number (there were 272 subjects), diet treatment group (there were 3), and weight loss (in kg) after 24 months

Subject	Group	WtLoss24
1	Low-Fat	2.2
2	Low-Fat	-4.8
3	Low-Fat	2.9
...		
105	Mediterranean	10.8
106	Mediterranean	6.4
107	Mediterranean	-0.3
...		
214	Low-Carbohydrate	3.4
215	Low-Carbohydrate	4.8
216	Low-Carbohydrate	10.9

DISPLAY 6.13

Volumes of INAH3 ($1,000 \times \text{mm}^3$) cell clusters from 41 human subjects at autopsy, by sex, sexual orientation, and cause of death

		Males		Females	
		Heterosexual	Homosexual	Heterosexual	
AIDS death	Non-AIDS death		AIDS death	AIDS death	Non-AIDS death
12	20	1	34	12	10
105	37	7	39		19
105	103	12	41		29
118	129	15	46		105
119	135	18	66		155
161	140	18	86		
	161	23	128		
	175	26	142		
	179	29	193		
	209	32			

cluster, INAH3, are re-created in Display 6.13. The numbers are 1,000 times volumes in mm^3 . Subjects are classified into five groups according to three factors: gender, sexual orientation, and cause of death. One male classified as a homosexual who died of AIDS (volume 128) was actually bisexual. LeVay used the term *presumed* heterosexual to indicate the possibility of misclassifying some subjects. Do heterosexual males tend to differ from homosexual males in the volume of INAH3? Do heterosexual males tend to differ from heterosexual females? Do heterosexual females tend to differ from homosexual males? Analyze the data and write a brief statistical report including a summary of statistical findings, a graphical display, and a details section describing the details of the particular methods used. Also describe the limitations of inferences that can be made. (Hint: What linear combination of the five means can be used to test whether cause of death can be ignored? If cause of death can be ignored, what linear combinations of the resulting three means are appropriate for addressing the questions above?)

Answers to Conceptual Exercises

1. (a) Yes, it is possible that the acting performance of the actor portrayed a more competent worker in the *crutches* role, even though the script was held constant. (b) With two pairs of actors and twice as many groups, the handicap effect could be isolated from the actor effect.
2. (a) (i) Hypothesis: all means are equal; alternative: at least one is different from the others. (ii) Hypothesis: all means are equal; alternative: the means are not equal but fall on a straight line function of male body size (with nonzero slope). (b) The alternative in (ii) is more specific. For a given set of data there is more power in detecting differences if the (correct) specific alternative can be investigated. (This has previously been noticed by the fact that a one-tailed p -value is smaller than a two-tailed.)
3. The researcher had reason to believe, because of other species of fish in the same genus, that the colored tail would be more attractive to the females.
4. If the new species average is somewhere between 1.6 and 3.8 Btu/lb, the range of the set of means is unchanged. If the new species average is either less than 1.6 or greater than 3.8 Btu/lb, the range is larger. Those are the only possibilities. So the answer is (d): the range cannot be smaller than the old range (but it could be larger). The range increases as the number of groups increase.
5. Where did the 65°F cutoff come from? If the analyst chose that cutoff because it produced the most dramatic difference between the two groups, the search procedure should be included in the assessment of evidence.
6. Yes, it does. It is important to pool information about variability because the population SD is difficult to estimate from small samples.
7. A planned comparison is one of a few specific comparisons that is designed to answer a question of interest. An unplanned comparison is one (of a large number) of comparisons that is suggested by the data themselves.
8. No. More complex comparisons can be made by examining linear combinations of group means.
9. The more groups there are, the larger the difference one expects between the smallest and largest averages. To incorporate the selection of this hypothesis on the basis of how the data turned out, the appropriate statistical measure of uncertainty is the same as the one that is appropriate for comparing every mean to every other mean.
10. Not necessarily. If there are no planned comparisons, it may be best to report no evidence of differences (protected LSD procedure). But the evidence about planned comparisons to answer questions of interest should be assessed on its own. It is possible that a planned comparison shows something when the F -test does not.
11. No. The parameter changes from γ to 3γ , the estimate changes from g to $3g$, and the standard error also changes from $SE(g)$ to $SE(3g) = 3SE(g)$. So the t -ratio is not changed at all. This is why one can take the convenient step of multiplying a set of coefficients by a common factor to make all coefficients into integers, if desirable.

Simple Linear Regression: A Model for the Mean

Chapters 5 and 6 advocated the investigation of specific questions of interest for the several-sample problem by paying attention to the structure of the grouped data. When the different groups correspond to different levels of a quantitative explanatory variable, the idea can be extended with the *simple linear regression* model, in which the means fall on a straight line function of the explanatory variable.

Such a model has several advantages when it is appropriate. It offers a concise summary of the mean of the response variable as a function of the explanatory variable through two parameters: the slope and the intercept of the line. For a surprisingly large number of data problems this model is appropriate (possibly after transformation) and the questions of interest can be conveniently reworded in terms of the two parameters. The parameters are estimable even without replicate values at each distinct level of the explanatory variable.

This chapter presents the model and some associated inferential tools. The next chapter takes a closer look at the simple linear regression model assumptions and how to measure departures from them.

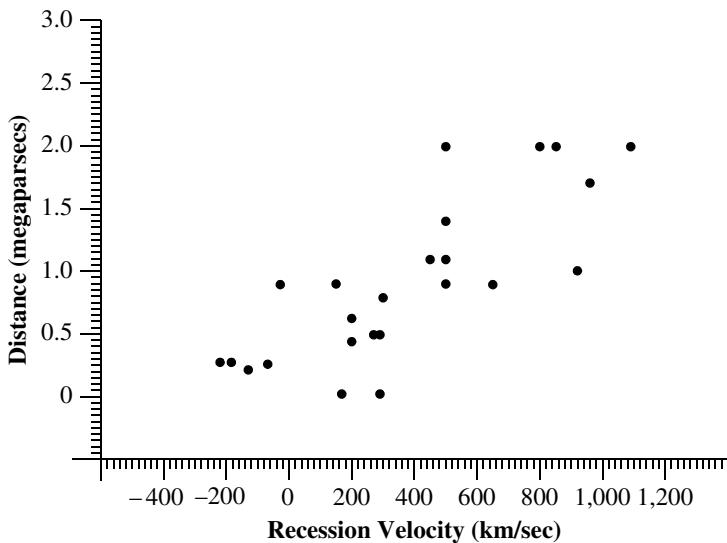
7.1 CASE STUDIES

7.1.1 The Big Bang—An Observational Study

Edwin Hubble used the power of the Mount Wilson Observatory telescopes to measure features of nebulae outside the Milky Way. He was surprised to find a relationship between a nebula's distance from earth and the velocity with which it was going away from the earth. Hubble's initial data on 24 nebulae are shown as a *scatterplot* in Display 7.1. (Data from E. Hubble, "A Relation Between Distance and Radial Velocity Among Extra-galactic Nebulae," *Proceedings of the National Academy of Science* 15 (1929): 168–73.) The horizontal axis measures the recession velocity, in kilometers per second, which was determined with considerable accuracy by the red shift in the spectrum of light from a nebula. The vertical scale measures distance from the earth, in megaparsecs: 1 megaparsec is 1 million parsecs, and 1 parsec is about 30.9 trillion kilometers. Distances were measured by comparing mean luminosities of the nebulae to those of certain star types, a method that is not particularly accurate. The data are shown later in this chapter as Display 7.8.

DISPLAY 7.1

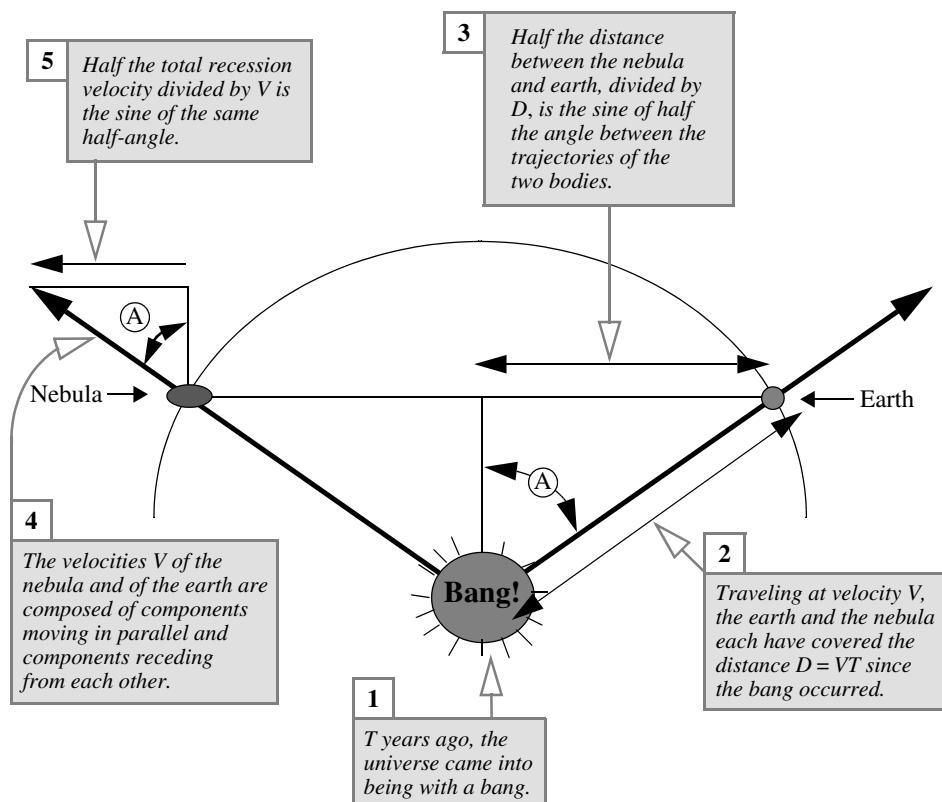
Scatterplot of measured distance versus velocity for 24 extra-galactic nebulae



The apparent statistical relationship between distance and velocity led scientists to consider how such a relationship could arise. It was proposed that the universe came into being with a Big Bang, a long time ago. The material in the universe traveled out from the point of the Big Bang, and scattered around the surface of an expanding sphere. If the material were traveling at a constant velocity (V) from the point of the bang, then the earth and any nebulae would appear as in Display 7.2.

DISPLAY 7.2

Big Bang theory model for distance–velocity relationship of nebulae



The distance (Y) between them and the velocity (X) at which they appear to be going away from each other satisfy the relationship

$$(Y/2)/VT = (X/2)/V = \sin(A),$$

where A is half the angle between them. In that case,

$$Y = TX$$

is a straight line relationship between distance and velocity. The points in Display 7.1 do not fall exactly on a straight line. It might be, however, that the *mean* of the distance measurements is TX . The slope parameter T in the equation $\text{Mean}\{Y\} = TX$ is the time elapsed since the Big Bang (that is, the age of the universe).

Several questions arise. Is the relationship between distance and velocity indeed a straight line? Is the y -intercept in the straight line equation zero, as the Big Bang theory predicts? How old is the universe?

Statistical Conclusion

If the theory is taken as correct, then the estimated age of the universe is 0.001922 megaparsecs-seconds per kilometer, or about 1.88 billion years (estimate of slope in simple linear regression through the origin). A 95% confidence interval for the age is 1.50 to 2.27 billion years. However, the data are not consistent with the Big Bang theory as proposed. Although the relationship between mean measured distance and velocity approximates a straight line, the value of the line at velocity zero is apparently not zero, as predicted by the theory (two-sided p -value = 0.0028 for a test that the intercept is zero).

Scope of Inference

These data are not a random sample, so they do not necessarily represent what would result from taking measurements from other nebulae. This analysis assumes that there is an exact linear relationship between distance and velocity, but that the measured distances do not fall exactly on a straight line because of measurement errors. The confidence interval above summarizes the uncertainty that comes from errors in distance measurements. Uncertainty due to errors in measuring velocities is not included in the p -values and confidence coefficients. Such errors are a potential source of pronounced bias. (*Note:* The Big Bang theory is still intact. Astronomers now estimate the universe to be 13.7 billion years old.)

7.1.2 Meat Processing and pH—A Randomized Experiment

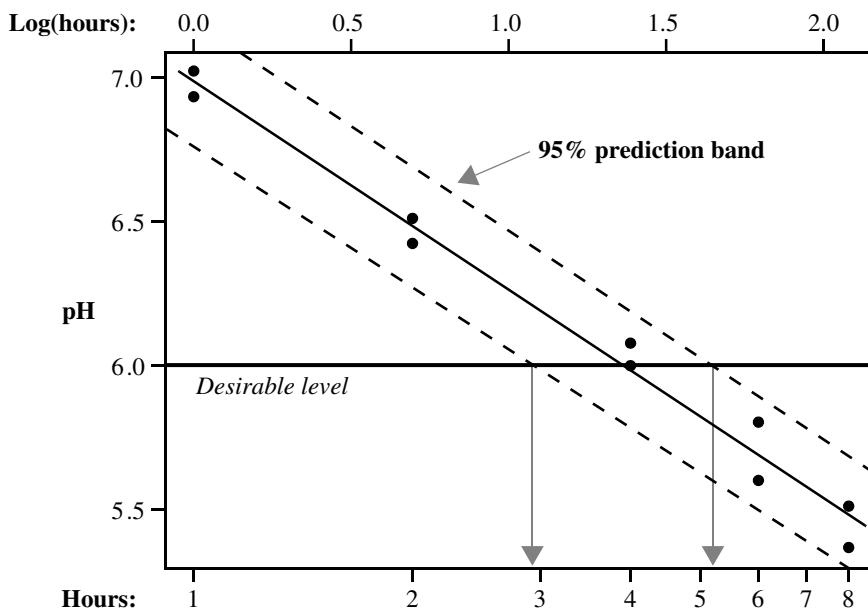
A certain kind of meat processing may begin once the pH in postmortem muscle of a steer carcass decreases to 6.0, from a pH at time of slaughter of around 7.0 to 7.2. It is not practical to monitor the pH decline for each animal, so an estimate is needed of the time after slaughter at which the pH reaches 6.0. To estimate this time, 10 steer carcasses were assigned to be measured for pH at one of five times after slaughter. The data appear in Display 7.3. (Data from J. R. Schwenke and

DISPLAY 7.3 pH of 10 steer carcasses measured at five different times after slaughter

Steer	Time after slaughter (hours)	pH
1	1	7.02
2	1	6.93
3	2	6.42
4	2	6.51
5	4	6.07
6	4	5.99
7	6	5.59
8	6	5.80
9	8	5.51
10	8	5.36

DISPLAY 7.4

Meat processing data with estimated regression line (from the simple linear regression of pH on log time after slaughter) and a 95% prediction band



G. A. Milliken, “On the Calibration Problem Extended to Nonlinear Models,” *Biometrics* 47 (1991): 563–74.)

Statistical Conclusion

The estimated relationship between postmortem muscle pH and time after slaughter is summarized in Display 7.4. The solid line shows the estimated mean pH as a function of the logarithm of time after slaughter. The dotted lines are the upper and lower endpoints of 95% prediction intervals for the pH of steer carcasses at times after slaughter in the range of 1 to 8 hours. It is estimated that the mean pH at 3.9 hours is 6. It is predicted that at least 95% of steer carcasses will reach a pH of 6.0 sometime between 2.94 and 5.10 hours after slaughter (95% calibration interval).

7.2 THE SIMPLE LINEAR REGRESSION MODEL

7.2.1 Regression Terminology

Regression analysis is used to describe the distribution of values of one variable, the *response*, as a function of other—*explanatory*—variables. This chapter deals with *simple* regression, where there is a single explanatory variable.

Think of a series of subpopulations of responses, one for each value of the explanatory variable, such as the subpopulation of pH values of steer carcasses, 4 hours after slaughter. The *regression of the response variable on the explanatory variable* is a mathematical relationship between the *means* of these subpopulations and the explanatory variable. The *simple linear regression* model specifies that this relationship is a straight line function of the explanatory variable. As evident in Display 7.4, a useful model for the meat processing data has the means of the subpopulations of pH falling on a straight line function of the logarithm of time after slaughter.

Let Y and X denote, respectively, the response variable and the explanatory variable. The notation $\mu\{Y|X\}$ will represent the regression of Y on X , and it should be read as “the mean of Y as a function of X .” When a specific value, $X = x$, is given to the explanatory variable, the same expression should be read as “the mean of Y when $X = x$.” The simple linear regression model is

$$\mu\{Y|X\} = \beta_0 + \beta_1 X.$$

The equation involves two statistical parameters. β_0 is the *intercept* of the line, measured in the same units as the response variable. β_1 is the *slope* of the line, equal to the rate of change in the mean response per one-unit increase in the explanatory variable. The units of β_1 are the ratio of the units of the response variable to the units of the explanatory variable.

A simple linear regression model for the statistical relationship between Y = Distance and X = Velocity in the Big Bang example is:

$$\mu\{\text{Distance}|\text{Velocity}\} = \beta_0 + \beta_1 \text{Velocity},$$

in which β_0 and β_1 are the intercept and slope of the regression line. The intercept is in units of megaparsecs and the slope is in units of megaparsecs per (km/sec).

A similar notation is $\sigma\{Y|X\}$ for the standard deviation of Y as a function of X . The equal-SD assumption specifies that $\sigma\{Y|X\} = \sigma$, for all X .

The complete picture of a probability model emerges if, as depicted in Display 7.5, each of the subpopulations has a normal distribution. This model has only three unknown parameters— β_0 , β_1 , and σ . Notice that the model describes the distributions of responses, but it says nothing about the distribution of the explanatory variable. In various applications, X -values may be chosen by the researcher, or they may themselves be random.

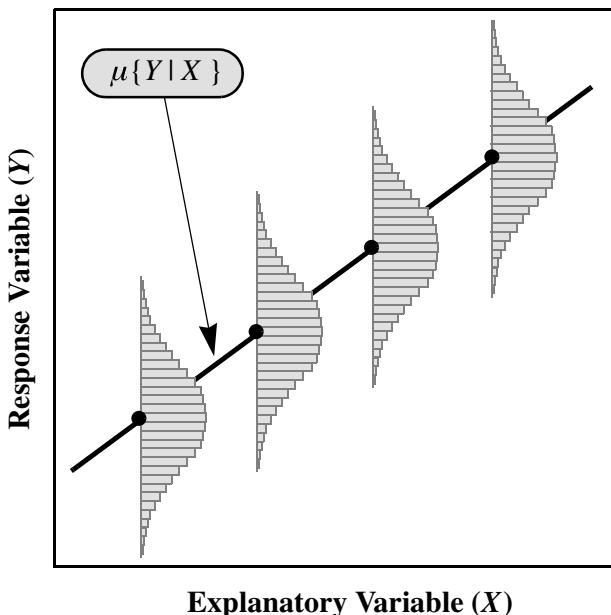
The independence assumption means that each response is drawn independently of all other responses from the same subpopulation and independently of all responses drawn from other subpopulations. The common types of violations are cluster effects and serial effects, as before.

7.2.2 Interpolation and Extrapolation

The simple linear regression model makes it possible to draw inference about any mean response within the range of the explanatory variable. Statements about the

DISPLAY 7.5

The ideal normal, simple linear regression model

***Model Assumptions***

1. There is a normally distributed subpopulation of responses for each value of the explanatory variable.
2. The means of the subpopulations fall on a straight line function of the explanatory variable.
3. The subpopulation standard deviations are all equal (to σ).
4. The selection of an observation from any of the subpopulations is independent of the selection of any other observation.

mean at values of X not in the data set, but within the range of the observed explanatory variable values, are called *interpolations*. The ability to interpolate is a strong advantage to regression analysis. Making statements for values outside of this range—*extrapolation*—is potentially dangerous, however, because the straight line model is not necessarily valid over a wider range of explanatory variable values.

The regression of a response variable on an explanatory variable is not necessarily a straight line. The simple regression model is very useful, though, because the regression is often well approximated by a straight line in the region of interest or is approximated by a straight line after transformation.

The remainder of this chapter assumes the simple regression model is appropriate for the case studies. Model checking, transformations, and alternatives will be discussed in the next chapter.

7.3 LEAST SQUARES REGRESSION ESTIMATION

The method of least squares is one of many procedures for choosing estimates of parameters in a statistical model. The method was described by Legendre in 1805, but the German mathematician Karl Gauss had been using it since 1795 and is generally regarded as its originator. The method provides the foundation for nearly all regression and analysis of variance procedures in modern statistical computer software.

7.3.1 Fitted Values and Residuals

The hat notation ($\hat{\cdot}$) denotes an estimate of the parameter under the hat, so $\hat{\beta}_0$ and $\hat{\beta}_1$ are symbols for estimates of β_0 and β_1 , respectively. Once these estimates are determined, they combine to form an estimated mean function,

$$\hat{\mu}\{Y|X\} = \hat{\beta}_0 + \hat{\beta}_1 X.$$

This general expression may be calculated for each X_i in the data set to estimate the mean of the distribution from which the corresponding response Y_i arose. The estimated mean is called the *fitted value* (or sometimes *predicted value*) and is represented by fit_i . The difference between the observed response and its estimated mean is the *residual*, denoted by res_i :

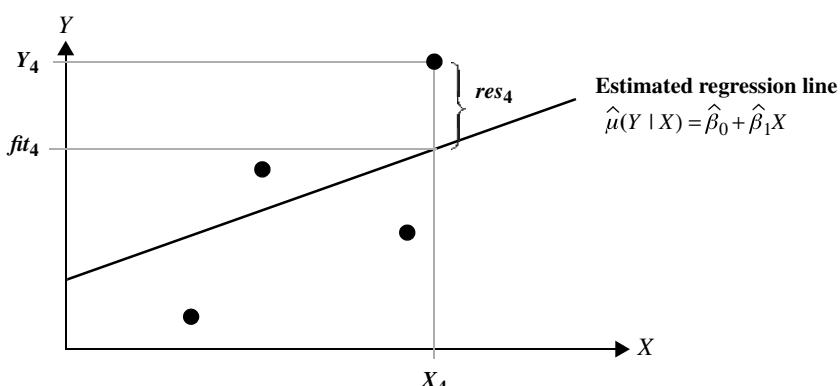
$$\text{fit}_i = \hat{\mu}\{Y_i|X_i\} = \hat{\beta}_0 + \hat{\beta}_1 X_i.$$

$$\text{res}_i = Y_i - \text{fit}_i$$

These definitions are illustrated in Display 7.6 for a simple example.

DISPLAY 7.6

Illustration of the residual and fitted value for observation (X_4, Y_4) in a hypothetical data set of size 4



The magnitude of a single residual measures the distance between the corresponding response and its fitted value according to the estimated mean function. A good estimate should result in a small distance. A measure of the distance between *all* responses and their fitted values is provided by the *residual sum of squares*.

7.3.2 Least Squares Estimators

Least squares estimates of β_0 and β_1 are those values for the intercept and slope that minimize the sum of squared residuals (hence “least squares”). The solution to this fitting criterion is found by calculus to be:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

(the interested reader is referred to Exercise 26). Statistical computer packages carry out the arithmetic, as shown for the Big Bang data in Display 7.9.

7.3.3 Sampling Distributions of the Least Squares Estimators

The sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ are the distributions of least squares estimates from hypothetical repeated sampling of response measurements at the same (fixed) values of the explanatory variable. In the meat processing example, a single repetition consists of obtaining pH readings from two new steers at each of the times 1, 2, 4, 6, and 8 hours after slaughter. If the normal, simple linear regression model applies, then Display 7.7 summarizes what statistical theory concludes about the sampling distribution.

7.3.4 Estimation of σ from Residuals

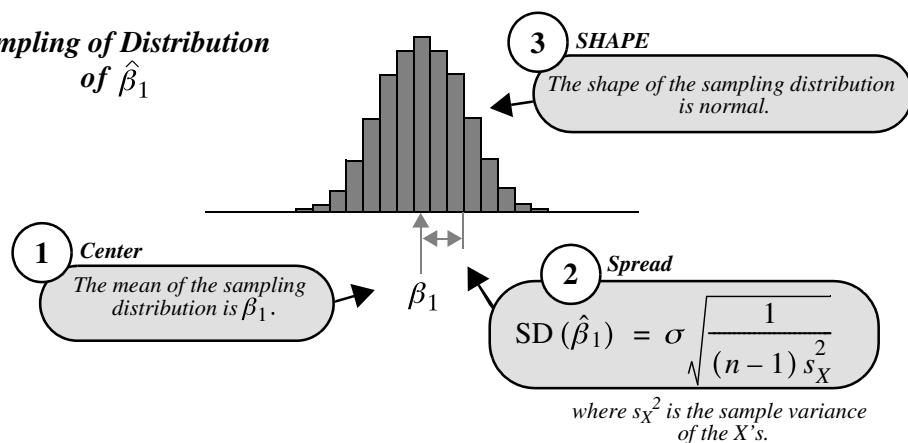
The parameters of primary interest are β_0 and β_1 , but the sampling distributions of their estimators depend on the parameter σ . Therefore, even if there is no direct interest in the variation of the observations about the regression line, an estimate of σ must be obtained in order to assess the uncertainty about estimates of the slope and the intercept. Residuals provide the basis for an estimate:

$$\hat{\sigma} = \sqrt{\frac{\text{Sum of all squared residuals}}{\text{Degrees of freedom}}},$$

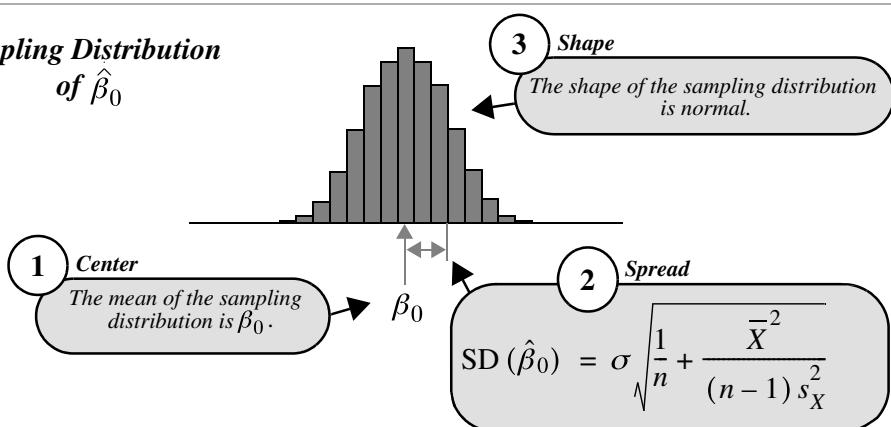
DISPLAY 7.7

Facts about the sampling distributions of the least squares estimators of slope and intercept in the ideal normal model (from statistical theory)

**Sampling of Distribution
of $\hat{\beta}_1$**



**Sampling Distribution
of $\hat{\beta}_0$**



where the degrees of freedom are the sample size minus *two*, $n - 2$. This is another instance of applying the general rule for finding the degrees of freedom associated with residuals from estimating a model for the means.

Rule for Determining Degrees of Freedom

*Degrees of freedom associated with a set of residuals =
(Number of observations) –
(Number of parameters in the model for the means).*

To understand the rule's application to simple linear regression, it is helpful to think about the case where $n = 2$. The least squares estimated line connects the two

points, and both residuals are zero. There are no degrees of freedom for estimating variance. Only by adding a third or more observations is there any information about the variability.

Example—Big Bang

The parameter estimates for the Hubble data are: $\hat{\beta}_0 = 0.3991$ megaparsecs and $\hat{\beta}_1 = 0.001373$ megaparsecs/(km/sec). The fitted values and residuals from the least squares fit are shown in Display 7.8. The sum of the 24 squared residuals is 3.6085, with 22 degrees of freedom. The estimate of σ^2 is 0.164, and the estimate of σ is 0.405 megaparsecs.

DISPLAY 7.8

Fitted values and residuals for the Big Bang study

<i>i</i>	Nebula	Velocity		Distance	
		<i>X_i</i>	<i>Y_i</i>	<i>fit_i</i>	<i>res_i</i>
1	S. Mag.	170	0.032	0.632	-0.600
2	L. Mag.	290	0.034	0.797	-0.763
3	NGC 6822	-130	0.214	0.221	-0.007
4	NGC 598	-70	0.263	0.303	-0.040
5	NGC 221	-185	0.275	0.145	0.130
6	NGC 224	-220	0.275	0.097	0.178
7	NGC 5457	200	0.450	0.674	-0.224
8	NGC 4736	290	0.500	0.797	-0.297
9	NGC 5194	270	0.500	0.770	-0.270
10	NGC 4449	200	0.630	0.674	-0.044
11	NGC 4214	300	0.800	0.811	-0.011
12	NGC 3031	-30	0.900	0.358	0.542
13	NGC 3627	650	0.900	1.292	-0.392
14	NGC 4626	150	0.900	0.605	0.295
15	NGC 5236	500	0.900	1.086	-0.186
16	NGC 1068	920	1.000	1.662	-0.662
17	NGC 5055	450	1.100	1.017	0.083
18	NGC 7331	500	1.100	1.086	0.014
19	NGC 4258	500	1.400	1.086	0.314
20	NGC 4151	960	1.700	1.717	-0.017
21	NGC 4382	500	2.000	1.086	0.914
22	NGC 4472	850	2.000	1.566	0.434
23	NGC 4486	800	2.000	1.497	0.503
24	NGC 4649	1090	2.000	1.896	0.104

7.3.5 Standard Errors

Standard errors for the intercept and slope estimates are obtained by replacing σ with $\hat{\sigma}$ in the formulas for their standard deviations (see Display 7.7). The degrees of freedom assigned to these standard errors are $n - 2$, the same as those attached to the residual estimate of the standard deviation:

$$\text{SE}(\hat{\beta}_1) = \hat{\sigma} \sqrt{\frac{1}{(n-1)s_X^2}}, \quad \text{d.f.} = n-2$$

and

$$\text{SE}(\hat{\beta}_0) = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_X^2}}, \quad \text{d.f.} = n-2,$$

where s_X^2 is the sample variance of the X 's.

Example—Big Bang

Display 7.9 shows a table that is typical of computer output for regression. Separate rows show the estimates, standard errors, and t -tests for the hypotheses that parameter values are zero. The first row corresponds to β_0 , the coefficient of the oxymoronic *constant variable*. The second row corresponds to β_1 , the coefficient of the variable *velocity*.

Regression parameter estimates for the Big Bang study					
Variable	Coefficient	Standard error	t-statistic	p-value	
Constant	0.3991	0.1185	3.369	0.0028	
Velocity	0.001373	0.000227	6.036	0.0000045	
Estimate of $\sigma = 0.4050$ (22 d.f.)					
Ratios of coefficient estimates to their standard errors			For hypotheses that the coefficients = 0		

Another recommended way to report the results of regression analysis is to present the estimated equation with standard errors in parentheses below the corresponding parameter estimates:

$$\text{Estimated Mean Distance} = 0.3991 + 0.001373 \text{ (Velocity)} \\ (0.1185) \quad (0.000227)$$

$$\text{Estimated SD of Distances} = 0.405 \text{ megaparsec (22 d.f.)}$$

7.4 INFERENTIAL TOOLS

7.4.1 Tests and Confidence Intervals for Slope and Intercept

Inferences are based on the t -ratios, $(\hat{\beta}_0 - \beta_0)/SE(\hat{\beta}_0)$ and $(\hat{\beta}_1 - \beta_1)/SE(\hat{\beta}_1)$, which have t -distributions with $n - 2$ degrees of freedom under the conditions of the normal simple linear regression model. As usual, *tests* are appropriate when the question is whether the data are consistent with some particular hypothesized value for the parameter, as in “Is 7 a possible value for β_0 ?” *Confidence intervals* are appropriate for presenting a range of values for a parameter that are consistent with the data, as in “What are the possible values of β_0 ?”

The p -values for the tests that β_0 and β_1 are individually equal to zero are included as standard output from statistical computer packages, as shown in Display 7.9. Although the test that β_1 is zero is often very important—so as to judge whether the mean of the response is unrelated to the explanatory variable—neither of the reported tests necessarily answers a relevant question. To obtain p -values for testing other hypothesized values or to construct confidence intervals, the user must perform other calculations based on the estimates and their standard errors. Some examples follow.

Does Hubble’s Data Support the Big Bang Theory? (Test That β_0 Is 0)

According to the theory discussed in Section 7.1.1, the true distance of the nebulae from earth is a constant times the recession velocity, so, if Y is measured distance, $\mu\{Y|X\} = \beta_1 X$. Thus, if the simple theory is right, the linear regression of measured distance on velocity will have intercept zero. Since the relationship does indeed appear to be a straight line, a check on the theory is accomplished by a test that β_0 is 0. From Display 7.9 the two-sided p -value for this test is reported as 0.0028, providing convincing evidence that these data do not support the theory as stated.

What Is the Age of the Universe According to the Big Bang Theory? (Confidence Interval for β_1)

According to the Big Bang theory, $\mu\{Y|X\} = \beta_1 X$, and β_1 is the age of the universe. To estimate β_1 with the value (0.001373) reported in Display 7.9 would be wrong, because that estimate refers to a different model—one with an intercept parameter—in which the slope parameter has a somewhat different meaning. A least squares fit to the Big Bang model—without the intercept parameter—gives an estimate for β_1 of 0.001922 megaparsec-second per km, with a standard error of 0.000191 based on 23 degrees of freedom. Since $t_{23}(0.975) = 2.069$, the confidence interval for β_1 is $0.001922 \pm (2.069)(0.000191)$, or from 0.001527 to 0.002317, or 1.50 to 2.27 billion years. The best estimate is 1.88 billion years. (One megaparsec-second per kilometer is about 979.8 billion years.)

7.4.2 Describing the Distribution of the Response Variable at Some Value of the Explanatory Variable

At some specified value, $X = X_0$, for the explanatory variable, the response distribution has mean $\beta_0 + \beta_1 X_0$ and standard deviation σ . For example, the distribution of pH levels of steers 4 hours after slaughter will have mean $\beta_0 + \beta_1 \log(4)$. Not all steers will have this pH; the variability of individual pH's around this mean is represented by the standard deviation σ . With least squares estimates, $\hat{\beta}_0 = 6.98$, $\hat{\beta}_1 = -0.7257$, and $\hat{\sigma} = 0.0823$, it can be estimated that the distribution of pH's at 4 hours has a mean of $6.98 - 0.7257 \log(4) = 5.98$ and an SD of 0.0823.

The Standard Error for the Estimated Mean

The standard error of $\hat{\mu} + \hat{\beta}_1 X_0$, for some specific number X_0 , is

$$\text{SE}[\hat{\mu}\{Y|X_0\}] = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{(n-1)s_X^2}}, \quad \text{d.f.} = n = 2.$$

Once this standard error is computed, a confidence interval or a test for the mean of Y at X_0 follows according to standard t -tools. In addition to $\hat{\sigma}$, the formula requires the average and the sample variance of the explanatory variable values in the data set. The calculations, as demonstrated on the meat processing example in Display 7.10, are straightforward, but inconvenient to do by hand.

DISPLAY 7.10

95% confidence interval for the mean pH of steers 4 hours after slaughter (from the estimated regression of pH on log(time) after slaughter for the meat processing data)

$$\begin{aligned} \hat{\mu}\{Y|1.386\} &= 6.9836 - 0.7257 \times 1.386 = 5.98 \\ \text{SE}[\hat{\mu}\{Y|1.386\}] &= 0.08226 \sqrt{\frac{1}{10} + \frac{(1.386 - 1.190)^2}{9(0.6344)}} \\ &= 0.0269 \\ t_{8(0.975)} &\\ \text{Upper limit: } 5.98 &+ 2.306 \times 0.0269 = 6.04 \\ \text{Lower limit: } 5.98 &- 2.306 \times 0.0269 = 5.92 \end{aligned}$$

Computer Centering Trick

A computer trick can be used to bypass the hand calculations. First, create an artificial explanatory variable, $X^* = X - X_0$, by subtracting X_0 from all explanatory variable values. This centers X at X_0 , in the sense that the value $X = X_0$ becomes the value $X^* = 0$. Next, fit the simple linear regression of Y on X^* . The *intercept* in the linear regression model $\mu\{Y|X^*\}$ is the mean of Y at $X^* = 0$, which is exactly the mean of Y at $X = X_0$. Therefore, you may now read the standard error for the estimated mean directly off the computer output as the standard error for the intercept parameter. This trick extends to many more complicated situations.

Additional Notes About the Estimated Mean at Some Value of X

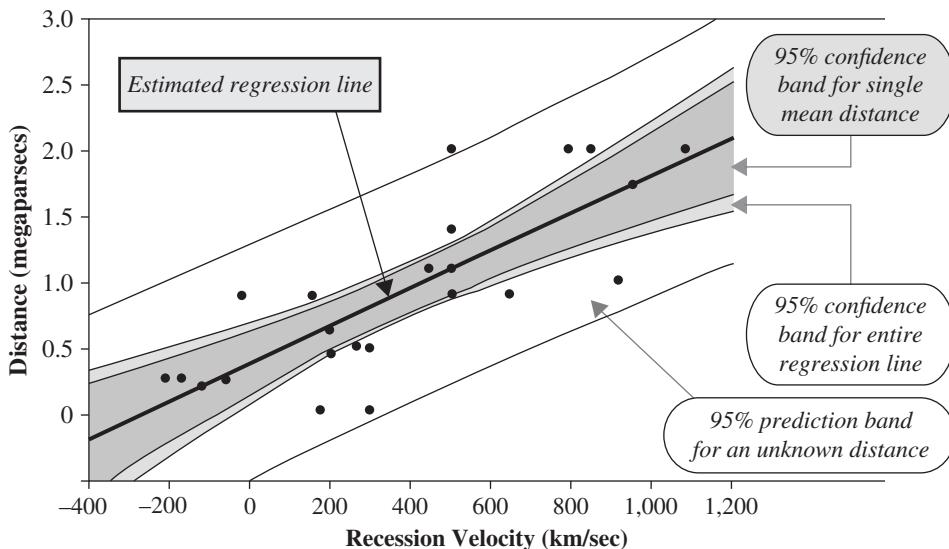
1. *SE for estimated intercept.* Notice that for $X_0 = 0$, the formula for the standard error of the estimated mean, above, reduces to the formula for the standard error of the estimated intercept.
2. *Precision in estimating $\mu\{Y|X\}$ is not constant for all values of X .* The formula for the standard error indicates that the precision of the estimated mean response decreases as X_0 gets farther from the sample average of X 's.
3. *Compound uncertainty in estimating several means simultaneously.* This section has addressed the situation where there is one single value of X_0 at which the mean response is desired. The construction of a 95% confidence interval is such that 95% of repetitions of the sampling process result in intervals that include the correct mean response *at that specified X_0* . But if one constructs different 95% confidence intervals for the mean responses at two values of X , the proportion of repetitions that result in both intervals correctly covering the respective means is something less than 95%. This is another case of compound uncertainty. To account for compound uncertainty in estimating mean responses at k different values of X , the Bonferroni procedure (Section 6.4.3) can be used. If k is very large, or if it is desired to obtain a procedure that protects against an unlimited number of comparisons, the Scheffé method can be used to construct a confidence band, as discussed next.
4. *Where is the regression line? (The Workman–Hotelling procedure)* Suppose the question is, “What is the mean response at all X -values within the observational range?” That is, where is the entire line, as opposed to where is one specific value on the line? By the device of replacing the t -multiplier in confidence intervals with a Scheffé multiplier based on an F -percentile, the confidence interval for the mean response can be converted to a *confidence band*, having the property that at least 95% of the repetitions mentioned above produce bands that include the correct mean response *everywhere*. If the 95th percentile in the F -distribution based on 2 and $(n - 2)$ degrees of freedom is $F_{2,n-2}(0.95)$, then the 95% confidence band has upper and lower bounds at each X given by

$$\hat{\mu}\{Y|X\} \pm \sqrt{2 \times F_{2,n-2}(0.95)} \times \text{SE}[\hat{\mu}\{Y|X\}].$$

A 95% confidence band for the regression line in the Hubble example uses $F_{2,22}(0.95) = 3.443$, so the multiplier for the standard error is 2.624 instead of the t -multiplier of 2.074. Band limits are calculated for different values of X and the confidence band is obtained by connecting all the upper endpoints and all lower endpoints, as shown in Display 7.11. Notice that the confidence band is not substantially wider than the intervals for individual means. The *prediction band* in the display is discussed in Section 7.4.3.

DISPLAY 7.11

The 95% confidence band on the population regression line, the 95% confidence interval band for single mean estimates, and a 95% prediction interval band for the Big Bang example



7.4.3 Prediction of a Future Response

Confidence intervals indicate likely values for parameters. Another inferential tool is a *prediction interval*, which indicates likely values for a future value of a response variable at some specified value of the explanatory variable.

Notice the difference between these two questions: (1) What is the *mean pH* 4 hours after slaughter? (2) What will be the *pH of a particular steer carcass* 4 hours after slaughter? The first question is about a parameter and the second is about an individual within the population. Even if the mean and standard deviation of pH at 4 hours after slaughter are known exactly, it is impossible to predict exactly the pH of an individual steer.

The best single prediction of a future response at X_0 , denoted by $\text{Pred}\{Y|X_0\}$, is the estimated mean response at X_0 :

$$\text{Pred}\{Y|X_0\} = \hat{\mu}\{Y|X_0\} = \hat{\beta}_0 + \hat{\beta}_1 X_0.$$

A *prediction interval* is an interval of likely values, along with a measure of the likelihood that the interval will include the future response value. This measure must account for two independent sources of uncertainty: uncertainty about the location of the subpopulation mean, and uncertainty about where the future value will be in relation to its mean. The error in using the predictor above is

$$\begin{aligned}
 Y - \text{Pred}\{Y|X_0\} &= Y - \hat{\mu}\{Y|X_0\} \\
 &= [Y - \mu\{Y|X_0\}] - [\hat{\mu}\{Y|X_0\} - \mu\{Y|X_0\}]. \\
 &= \boxed{\text{Prediction error}} = \boxed{\text{Random sampling error}} + \boxed{\text{Estimation error}}
 \end{aligned}$$

The variance of the random sampling error, as evident from the definition of the simple linear regression model, is σ^2 . The variance of the estimation error is the variance of the sampling distribution of $\hat{\beta}_0 + \hat{\beta}_1 X_0$. The standard error of prediction combines the estimated variances as

$$\text{SE}[\text{Pred}\{Y|X_0\}] = \sqrt{\hat{\sigma}^2 + \text{SE}[\hat{\mu}\{Y|X_0\}]^2}.$$

Some statistical programs provide standard errors of prediction at requested values. If not, both quantities under the square root sign will be available from computer output, if the computer centering trick of the previous section is invoked. As a last resort, the formula for hand calculation is

$$\text{SE}[\text{Pred}\{Y|X_0\}] = \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{(n-1)s_X^2}},$$

where, again, s_X^2 is the sample variance of the X 's.

Prediction intervals may be obtained from this standard error and a multiplier from the t -distribution with $n - 2$ degrees of freedom. Display 7.12 shows the construction of a 95% prediction interval for the pH's of steers at 4 hours after slaughter. The 95% prediction interval succeeds in capturing the future Y in 95% of its applications. The prediction bands in Displays 7.4 and 7.11 are the curves obtained by connecting the upper endpoints and connecting the lower endpoints of prediction intervals at many different X -values.

7.4.4 Calibration: Estimating the X That Results in $Y = Y_0$

Sometimes it is desired to estimate the X that produces a specific response Y . This happens to be the case for the meat processing example: "At what time after slaughter will the pH in any particular steer be 6?" This is a *calibration* problem, also known as *inverse prediction*.

DISPLAY 7.12

95% prediction interval for the pH of a steer carcass 4 hours after slaughter (from the estimated regression of pH on log time after slaughter)

$$\begin{aligned} \text{Pred}\{Y|1.386\} &= 6.9836 - 0.7257 \times 1.386 = 5.98 \\ \text{SE}[\text{Pred}\{Y|1.386\}] &= 0.08226 \sqrt{1 + \frac{1}{10} + \frac{(1.386 - 1.190)^2}{9(0.6344)}} \\ &= 0.0865 \\ \text{Upper limit: } &5.98 + 2.306 \times 0.0865 = 6.18 \\ \text{Lower limit: } &5.98 - 2.306 \times 0.0865 = 5.78 \end{aligned}$$

In a typical calibration problem, an experiment is used to estimate the regression of an imprecise measurement (Y) of some quantity on its actual value (X), as determined by a more precise, but more expensive measuring instrument. The regression provides a calibration equation, so that the accurate measurement can be predicted from the cheaper, inaccurate one. It may seem appropriate to call the precise value Y , to use the imprecise value as X , and to employ prediction as in the previous section. This is not possible, however, when the precise values are controlled by the researcher, because the distributional model makes no sense.

The simplest method for calibration is to invert the prediction relationship. If $\text{Pred}\{Y|X_0\}$ is taken to be $\hat{\beta}_0 + \hat{\beta}_1 X_0$, then $\text{Pred}\{X|Y_0\}$ is $(Y_0 - \hat{\beta}_0)/\hat{\beta}_1$. This is the value of X at which the predicted Y is Y_0 . A graphical method for obtaining a calibration interval (inverse prediction interval) for the X at some Y_0 is to plot prediction bands (the ones for predicting Y from X), draw a horizontal line at Y_0 , then draw vertical lines down to the X -axis from the points where the horizontal line intersects the upper and lower prediction limits. These two points are the limits of the calibration interval. This procedure is shown for the meat processing data in Display 7.4.

The uncertainty in a calibration interval has to do with the prediction of new Y 's. For the meat processing data: at 2.94 hours after slaughter (i.e., when \log hours = 1.08), it is predicted that 95% of steer carcasses will have a pH between 6.0 and 6.3. At 5.10 hours (when \log hours = 1.63) it is predicted that 95% of steer carcasses will have a pH between 5.7 and 6.0. According to these statements, only 2.5% of carcasses will have a pH less than 6.0 at time 2.94 hours, and only 2.5% of carcasses will have a pH above 6.0 at time 5.10 hours.

A similar method directly estimates (or predicts) $\hat{X} = (Y_0 - \hat{\beta}_0)/\hat{\beta}_1$ with an approximate standard error of either

$$\text{SE}(\hat{X}) = \frac{\text{SE}(\hat{\mu}\{Y|\hat{X}\})}{|\hat{\beta}_1|}$$

(to estimate the X at which the mean of Y is Y_0) or

$$\text{SE}(\hat{X}) = \frac{\text{SE}(\text{Pred}\{Y|\hat{X}\})}{|\hat{\beta}_1|}$$

(to find the X at which the value of Y is Y_0). An interval is centered at \hat{X} , with a half-width equal to a t -multiplier (d.f. = $n - 2$) times the standard error. (*Note:* When the regression line is too flat, these methods are not satisfactory.)

Example—Meat Processing

The log time at which the mean pH has decreased to 6.0 is estimated to be $\hat{X} = 1.3554$. Its standard error is 0.0367, so a 95% confidence interval is from 1.2708 to 1.4400. Hence the estimate of time required is 3.88 hours, with an approximate 95% confidence interval from 3.56 to 4.22 hours. To predict the time when the pH of a single steer might decline to 6.0, the standard error must account for the pH not being at its average. The second standard error is 0.1191, so the interval for $\log(\hat{X})$ is from 1.0806 to 1.6301. The estimate of the time is the same, 3.88 hours, but the approximate 95% confidence interval is from 2.95 to 5.10 hours. The interval is similar to the one determined graphically in Display 7.4.

7.5 RELATED ISSUES

7.5.1 Historical Notes About Regression

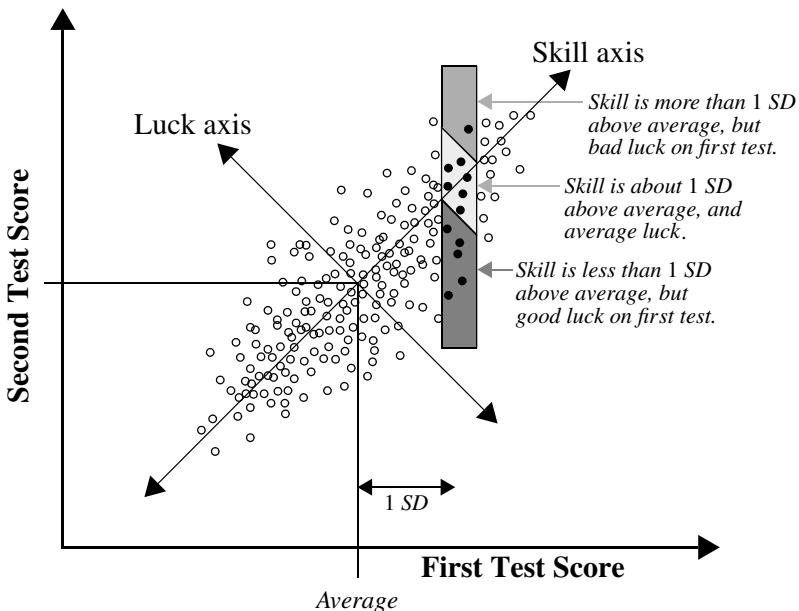
Regression analysis was first used by Sir Francis Galton, a 19th-century English scientist studying genetic similarities in parents and offspring. In describing the mean of child height as a function of parent height, Galton found that children of tall parents tended to be tall, but on average not as tall as their parents. Similarly, children of short parents tended to be short, but on average not as short as their parents. He described this as “regression towards mediocrity.” Others using his method to analyze statistical relationships referred to the method as Galton’s “regression line,” and the term *regression* stuck. (See Exercise 7.26.)

The Regression Effect

Galton’s phenomenon of regression toward the mean appears in any test-retest situation. Subjects who score high on the first test will, as a group, score closer to the average (lower) on the second test, while subjects who score low on the first test will, as a group, also score closer to the average (higher) on the second test. The

DISPLAY 7.13

Test-retest scores, illustrating the regression effect



reason is illustrated in Display 7.13. Each point shows the scores of a single subject on the two tests. The horizontal axis contains the first test score, and a vertical box isolates a set of subjects who scored about one standard deviation above average on the first test. Subjects in the set are there (1) because their overall skill is about 1 SD above average, and they performed at their skill level; (2) because their overall skill level was lower than 1 SD above average, and they performed somewhat above their overall skill level on this test; or (3) because their overall skill level was higher than 1 SD above average, and they performed somewhat below their overall skill level on this test. Performances on single tests that differ from one's overall skill level might be called luck or chance.

As is apparent in Display 7.13, more subjects fall into category (2) than into category (3), and they bring the average of the second test scores to a level lower than 1 SD above average. This is the *regression effect*, and it is real. The reason for the regression effect is that there are many more individuals in the whole set with skill levels near average than there are with skill levels farther than one standard deviation from it. So, by chance, more appear in the strip whose skill level is closer to the average (and their skill level tends to catch up with them on the second test).

The regression effect is therefore a natural phenomenon in any test-retest situation. It is often mistakenly interpreted as something more, suggesting some theory about the attitude toward the second test of the test takers based on their first test scores. The term *regression fallacy* is used to describe the mistake of attaching some special meaning to the regression effect.

7.5.2 Differing Terminology

The terms *dependent variable* and *independent variable* have been used in the past for response and explanatory variables, respectively. *The Statistical Sleuth* uses *response* and *explanatory* to avoid confusion with the notion of statistical independence, which is entirely different. (Economists often use *endogenous* and *exogenous*, respectively, to distinguish between variables that are determined within the system under study and those determined externally to it. However, the merits of this terminology have not led to its adoption by a wider circle of users.)

7.5.3 Causation

There is a tendency to think of the explanatory variable as a causative agent and the response as an effect, even with observational studies. But the same conditions apply here as before. With randomized experiments, cause-and-effect terminology is appropriate; with observational studies, it is not. With data from observational studies, the most satisfactory terminology describes an *association* between the mean response and the value of the explanatory variable.

7.5.4 Correlation

The *sample correlation coefficient* for describing the degree of linear association between any two variables X and Y is

$$r_{XY} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})/(n - 1)}{s_X s_Y},$$

where s_X and s_Y are the sample standard deviations. Unlike the estimated slope in regression, the sample correlation coefficient is symmetric in X and Y , so it does not depend on labeling one of them the response and one of them the explanatory variable.

A correlation is dimension-free. It falls between -1 and $+1$, with the extremes corresponding to the situations where all the points in a scatterplot fall exactly on a straight line with negative and positive slopes, respectively. Correlation of zero corresponds to the situation where there is no linear association.

Although the sample correlation coefficient is always a useful summary measure, inferences based on it only make sense when the pairs (X, Y) are selected randomly from a population. This is more restrictive than the conditions for inference from the regression model, where X need not be random. It is a common mistake to base conclusions on correlations when the X 's are not random.

If the data are randomly selected pairs, then the sample correlation coefficient estimates a population correlation, $\rho_{XY} = \text{Mean}\{(X - \mu_X)(Y - \mu_Y)\}/(\sigma_X \sigma_Y)$. Conveniently, a test of the hypothesis that $\rho_{XY} = 0$ turns out to be identical to the test of the hypothesis that the slope in the linear regression of Y on X is zero.

Tests for nonzero hypothesized values are also available, but they are quite different and are not pursued here.

Correlation only measures the degree of *linear* association. It is possible for there to be an exact relationship between X and Y and yet the sample correlation coefficient is zero. This would be the case, for example, if there was a parabolic relationship with no linear component.

7.5.5 Planning an Experiment: Replication

Replication means applying exactly the same treatment to more than one experimental unit. With the 10 steers available for the meat processing experiment, the researchers decided to administer five treatment levels—the five different times after slaughter—with two replicate observations at each level. Although they could have investigated 10 different times after slaughter, the benefits of replication can outweigh the extra information that would come from more explanatory variable values. The main benefit is that replication allows for an estimate of σ^2 that does not depend on any model being correct (the pooled estimate of the sample variances). This is sometimes referred to as the *pure error* estimate of σ^2 . This, in turn, permits an assessment of model adequacy that would not otherwise be available. (The details of a formal test for model adequacy are shown in Section 8.5.3.) A secondary benefit is a clearer picture of the relationship between the variance and the mean, which may permit an easier assessment of the need for and the type of transformation for the response variable.

Even though the benefits of replication are many, the practical decision about allocating a fixed number of experimental units to treatment levels is difficult. The researcher must weigh the benefits of replication against the benefits of further treatment levels. It would not be good in the meat processing study, for example, to place all 10 steers in only two treatment levels unless the researchers were certain that the regression was a straight line.

7.6 SUMMARY

In the simple linear regression model the mean of a response variable is a straight line function of an explanatory variable. The slope and intercept of the regression line can be estimated by the method of least squares. The variance about the regression is estimated by the sum of squared residuals divided by the degrees of freedom $n - 2$. Inferential tools discussed in this chapter are t -tests and confidence intervals for slope and intercept, t -tests and confidence intervals for the mean of the response at any particular value of the explanatory variable, prediction intervals for a future response at any particular value of the explanatory variable, and calibration intervals for predicting the X associated with a specified Y .

Big Bang Analysis

The analysis begins with a scatterplot of distance versus velocity. It is apparent from this (Display 7.1) that the regression of distance on velocity is at least approximately

a straight line, as predicted by the Big Bang theory. The least squares method provides a way to estimate the slope and intercept in this regression. Two questions are asked: is the intercept zero, as predicted by the theory, and what is the slope; which, according to the theory, is the age of the universe? The evidence in answer to the first question is expressed through a *p*-value for the test that β_0 is zero. The evidence in answer to the second is expressed through a confidence interval for β_1 .

Meat Processing Analysis

Even though the question of interest calls for finding the value of time after slaughter at which pH is 6, it is necessary to specify pH to be the response and time (log hours after slaughter) to be the explanatory variable, since the latter was controlled by the researchers. The regression of pH on log time is well approximated by a straight line, at least for times up to 8 hours. The calibration-like question of interest requires the data analyst to find the value of X (log time) at which Y (pH) is predicted to be 6.

7.7 EXERCISES

Conceptual Exercises

- 1. Big Bang Data.** Can the estimated regression equation be used to make inferences about (a) the mean distance of nebulae whose recession velocities are zero? (b) the mean distance of nebulae whose recession velocities are 1,000 km/sec? (c) the mean distance of nebulae whose velocities are 2,000 km/sec? (See Display 7.1.)
- 2. Big Bang Data.** Explain why improved measurement of distance would lead to more precise estimates of the regression coefficients.
- 3. Meat Processing Data.** By inspecting Display 7.4, describe the distribution of pH's for steer carcasses 1.65 hours after slaughter (where $X = \log(1.65) = 0.5$).
- 4.** What is wrong with this formulation of the regression model: $Y = \beta_0 + \beta_1 X$?
- 5.** Consider a regression of weight (kg) on height (cm) for a sample of adult males. What are the units of measurement of (a) the intercept? (b) the slope? (c) the SD? (d) the correlation?
- 6.** Explain the differences between the following terms: regression, regression model, and simple linear regression model.
- 7.** What assumptions are made about the distribution of the explanatory variable in the normal simple linear regression model?
- 8.** A group of children were all given a dexterity test in their fifth grade physical education class. The teacher noted a small group who performed exceptionally well, and she informed the grade six teacher as to which children were in the group. When the same children were given a similar dexterity test the next year, they performed reasonably well, but not as well as the sixth grade teacher had expected. What might have caused this?
- 9.** Consider the regression of weight on height for a sample of adult males. Suppose the intercept is 5 kg. (a) Does this imply that males of height 0 weigh 5 kg, on average? (b) Would this imply that the simple linear regression model is meaningless?

- 10.** (a) At what value of X will there be the most precise estimate of the mean of Y ? (b) At what value of X will there be the most precise prediction of a future Y ?
- 11.** What is the standard error of prediction as the sample size approaches infinity?

Computational Exercises

- 12.** Suppose that the fit to the simple linear regression of Y on X from 6 observations produces the following residuals: $-3.3, 2.1, -4.0, -1.5, 5.1, 1.6$. (a) What is the estimate of σ^2 ? (b) What is the estimate of σ ? (c) What are the degrees of freedom?
- 13. Big Bang Data.** Using the results shown in Display 7.9, find a 95% confidence interval for the intercept in the regression of measured distance on recession velocity.
- 14. Big Bang Data.** Using the results in Display 7.9, find a 95% confidence interval for the slope in the regression of measured distance on recession velocity.
- 15. Pollen Removal.** Reconsider the data in Exercise 3.27. (a) Draw a scatterplot of proportion of pollen removed versus duration of visit, for the bumblebee queens. (b) Fit the simple linear regression of proportion of pollen removed on duration of visit. Draw the estimated regression line on the scatterplot. (Do this with the computer, if possible; otherwise draw the line with pencil and ruler.) Is there any indication of a problem with this fit? (This problem will be continued in the next chapter.)
- 16.** Most computational procedures utilize the following identities:

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - \frac{1}{n} \left(\sum_{i=1}^n X_i \right)^2$$

and

$$\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^n X_i Y_i - \frac{1}{n} \left(\sum_{i=1}^n X_i \right) \left(\sum_{i=1}^n Y_i \right).$$

The left-hand expressions are needed for calculating $\hat{\beta}_1$, but the right-hand equivalents require less effort. Using the meat processing data, calculate the left-hand versions. Then calculate the right-hand versions. Did you get the same answers? Which version requires fewer steps? Which version requires less storage in a computer?

- 17. Meat Processing.** (a) Enter the data from Display 7.3 into a computer and find the least squares estimates for the simple linear regression of pH on log hours. (b) Noting that the average and sample standard deviation of X are provided in Display 7.12, calculate the estimated mean pH at 5 hours after slaughter, and the standard error of the estimated mean (using the formula in Section 7.4.2). (c) Find the standard error of the estimated mean pH at 5 hours after slaughter using the computer trick described in Section 7.4.2.
- 18. Meat Processing.** (a) Find the standard error of prediction for the prediction of pH at 5 hours after slaughter. (b) Construct a 95% prediction interval at 5 hours after slaughter.
- 19. Meat Processing.** Compute a 95% calibration interval (using the graphical approach) for the time at which pH of steer carcasses should be 7.
- 20. Meat Processing (Sample Size Determination).** The standard error of the estimated slope based on the 10 data points is 0.0344. Using the formula for SE in Section 7.3.5, and supposing that the spread of the X 's and the estimate of σ will be about the same in a future study, calculate how large the sample size would have to be in order for the SE of the estimated slope to be 0.01.
- 21. Planetary Distances and Order from the Sun.** The first three columns in Display 7.14 show the distances from the sun (scaled so that earth = 1) and the order from the sun of the 8 planets in

DISPLAY 7.14

Orders and distances from the sun (in astronomical units, so that the distance from earth to the sun is 1) of planets in our solar system; without and with the asteroid belt

Name	Order	Distance	Name 2	Order 2	Distance 2
Mercury	1	0.387	Mercury	1	0.387
Venus	2	0.723	Venus	2	0.723
Earth	3	1	Earth	3	1
Mars	4	1.524	Mars	4	1.524
Jupiter	5	5.203	(asteroids)	5	2.9
Saturn	6	9.546	Jupiter	6	5.203
Uranus	7	19.2	Saturn	7	9.546
Neptune	8	30.09	Uranus	8	19.2
Pluto	9	39.5	Neptune	9	30.09
			Pluto	10	39.5

our solar system and the dwarf planet, Pluto. The last three columns are the same but also include the asteroid belt. Using the first three columns, (a) draw a scatterplot of log of distance versus order, (b) include the least squares estimated simple linear regression line on the plot, (c) find the estimate of σ from the least squares fit, and (d) draw a scatterplot of the residuals versus the fitted values from this fit. Using the last three columns, (e) draw a scatterplot of log of distance versus order, (f) include the least squares estimated simple linear regression line on the plot, (g) find the estimate of σ from the least squares fit, and (h) draw a scatter plot of the residuals versus the fitted values from this fit. (i) Does it appear that the simple linear (straight line) regression model fits better to the first set of 9 planets or the second set of 10 “planets”? Explain.

22. Crab Claw Size and Force. As part of a study of the effects of predatory intertidal crab species on snail populations, researchers measured the mean closing forces and the propodus heights of the claws on several crabs of three species. Their data (read from their Figure 3) appear in Display 7.15. (Data from S. B. Yamada and E. G. Boulding, “Claw Morphology, Prey Size Selection and Foraging Efficiency in Generalist and Specialist Shell-Breaking Crabs,” *Journal of Experimental Marine Biology and Ecology*, 220 (1998): 191–211.)

- (a) Estimate the slope in the simple linear regression of log force on log height, separately for each crab species. Obtain the standard errors of the estimated slopes.
- (b) Use a *t*-test to compare the slopes for *C. productus* and *L. bellus*. Then compare the slopes for *C. productus* and *H. nudus*. The standard error for the difference in two slope estimates from independent samples is the following:

$$\text{SE}[\hat{\beta}_{1(1)} - \hat{\beta}_{1(2)}] = \sqrt{[\text{SE}(\hat{\beta}_{1(1)})]^2 + [\text{SE}(\hat{\beta}_{1(2)})]^2},$$

where $\hat{\beta}_{1(j)}$ represents the estimate of slope from sample j . Use *t*-tests with the sum of the degrees of freedom associated with the two standard errors. What do you conclude? (Note: A better way to perform this test, using multiple regression, is described in Chapter 9.)

23. For Those Literate in Calculus and Linear Algebra. The least squares problem is that of finding estimates of β_0 and β_1 that minimize the sum of squares,

$$\text{SS}(\beta_0, \beta_1) = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2.$$

DISPLAY 7.15 Closing force (in newtons) and propodus heights (in mm) in three predatory crab species

Hemigrapsus nudus (n = 14)		Lophopanopeus bellus (n = 12)		Cancer productus (n = 12)	
Force	Height	Force	Height	Force	Height
3.2	5.0	2.1	5.1	5.0	6.7
6.4	6.0	8.7	5.9	7.8	7.1
2.0	6.4	2.9	6.6	14.6	11.2
2.0	6.5	6.9	7.2	16.8	11.4
4.9	6.6	8.7	8.6	17.7	9.4
3.0	7.0	15.1	7.9	19.8	10.7
2.9	7.9	14.6	8.1	19.6	13.1
9.5	7.9	17.6	9.6	22.5	9.4
4.0	8.0	20.6	10.2	23.6	11.6
3.4	8.2	19.6	10.5	24.4	10.2
7.4	8.3	27.4	8.2	26.0	12.5
2.4	8.8	29.4	11.0	29.4	11.8
4.0	12.1				
5.2	12.2				

(a) Setting the partial derivatives of $\text{SS}(\beta_0, \beta_1)$ with respect to each parameter equal to zero, show that β_0 and β_1 must satisfy the *normal equations*:

$$\begin{aligned}\beta_0 n + \beta_1 \sum_{i=1}^n X_i &= \sum_{i=1}^n Y_i \\ \beta_0 \sum_{i=1}^n X_i + \beta_1 \sum_{i=1}^n X_i^2 &= \sum_{i=1}^n X_i Y_i.\end{aligned}$$

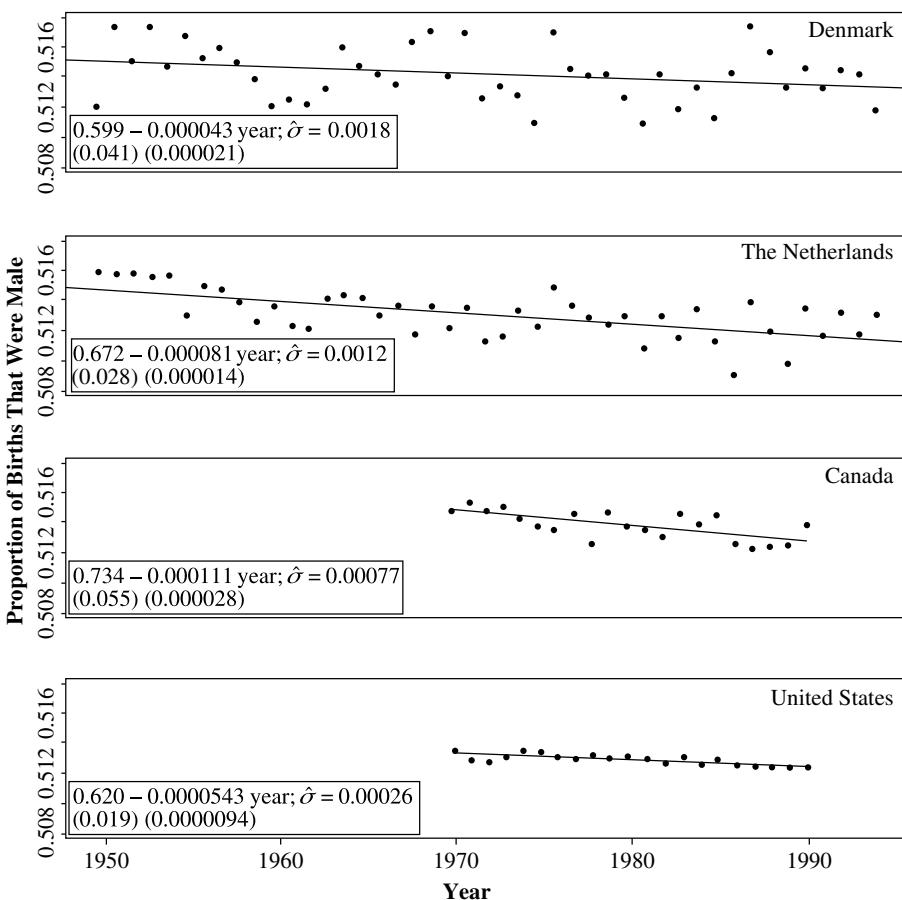
(b) Show that $\hat{\beta}_0$ and $\hat{\beta}_1$ given in Section 7.3.1 satisfy the normal equations. (c) Verify that the solutions provide a minimum to the sum of squares.

24. Decline in Male Births. Display 7.16 shows the proportion of male births in Denmark, The Netherlands, Canada, and the United States for a number of years. (Data read from graphs in Davis et al., “Reduced Ratio of Male to Female Births in Several Industrial Countries,” *Journal of the American Medical Association*, 279 (1998): 1018–23.) Also shown are the results of least squares fitting to the simple linear regression of proportion of males on year, separately for each country, with standard errors of estimated coefficients in parentheses.

- (a) With a statistical computer package and the data in the file ex0725, obtain the least squares fits to the four simple regressions, individually, to confirm the estimates and standard errors presented in Display 7.17.
- (b) Obtain the t -statistic for the test that the slopes of the regressions are zero, for each of the four countries. Is there evidence that the proportion of male births is truly declining?
- (c) Explain why the United States can have the largest of the four t -statistics (in absolute value) even though its slope is only the third largest (in absolute value).
- (d) Explain why the standard error of the estimated slope is smaller for the United States than for Canada, even though the sample size is the same.

DISPLAY 7.16

Proportions of male births in four countries; and simple regression statistics



- (e) Can you think of any reason why the standard deviations about the regression line might be different for the four countries? (*Hint:* The proportion of males is a kind of average, i.e., the average number of births that are male.)

Data Problems

- 25. Big Bang II.** The data in Display 7.17 are measured distances and recession velocities for 10 clusters of nebulae, much farther from earth than the nebulae reported in Section 7.1.1. (Data from E. Hubble and M. Humason, “The Velocity–Distance Relation Among Extra-Galactic Nebulae,” *Astrophysics Journal* 74 (1931): 43–50.) If Hubble’s theory is correct, then the mean of the measured distance, as a function of velocity, should be β_1 Velocity, and β_1 is the age of the universe. Are the data consistent with the theory (that the intercept is zero)? What is the estimated age of the universe? (*Note:* The slope here is in units of megaparsecs-seconds per kilometer. Multiply by 979.8 to get an answer in billions of years. You should find out how to fit simple linear regression through the

DISPLAY 7.17 Measured distance (million parsecs) and recession velocity (km/sec) for 10 clusters of nebulae

Cluster	Distance	Velocity
Virgo	1.8	890
Pegasus	7.25	3,810
Pisces	7.00	4,638
Cancer	9.00	4,820
Perseus	11.00	5,230
Coma	13.80	7,500
Ursa Major	22.00	11,800
Leo	32.00	19,600
Isolated nebulae I	4.20	2350
Isolated nebulae II	2.15	630

DISPLAY 7.18 Galton's data on heights (in inches) of adult children and their parents; first 5 of 933 rows

Gender	Family	Height	Father	Mother
Male	1	73.2	78.5	67
Female	1	69.2	78.5	67
Female	1	69	78.5	67
Female	1	69	78.5	67
Male	2	73.5	78.5	66.5

origin—that is, how to drop the intercept term—with your statistical computer package.) To what extent is the relationship shown by these far-away nebulae clusters similar to and different from the relationship indicated in Case Study 7.1.1? (Analyze the data and write a brief statistical report including a summary of statistical findings, a graphical display, and a details section describing the details of the particular methods used.)

26. Origin of the Term *Regression*. Motivated by the work of his cousin, Charles Darwin, the English scientist Francis Galton studied the degree to which human traits were passed from one generation to the next. In an 1885 study, he measured the heights of 933 adult children and their parents. Display 7.18 shows the first five rows of his data set. Galton multiplied all female heights by 1.08 to convert them to a male-equivalent height. He estimated the simple linear regression line of child's height on average parent's height and, upon finding that the slope was positive but less than 1, concluded that children of taller-than-average parents tended to also be taller than average but not as tall as their parents; and, similarly, children of shorter-than-average parents tended to be shorter than average, but not as short as their parents. He labeled this “regression towards mediocrity” because of the apparent regression (i.e., reversion) of children's height toward the average. Other scientists began to refer to the line as “Galton's regression line”. Although the term was not intended to describe the model for the mean, the name stuck. Reproduce Galton's analysis by converting females' heights to their male-equivalents (multiply them by 1.08). Do the same for mothers' heights. Compute the parent height by taking the average of the father's height and the converted mother's height. Then fit the simple linear regression of child height on parent height. (For now, do as Galton did and ignore the fact that the heights of individuals from the same family are probably not independent.) Find the predicted height and a 95% prediction interval for the adult height of a boy whose average parent height is 65 inches. Repeat for a boy whose average parent height is 76 inches. (These are the raw

data used in the paper by F. Galton, “Regression towards Mediocrity in Hereditary Stature” in the *Journal of the Anthropological Institute* in 1886, as described in “‘Transmuting’ Women into Men: Galton’s Family Data On Human Stature” by James Henley in *The American Statistician*, 58 (2004): 237–43.)

27. Male Displays. Black wheatears, *Oenanthe leucura*, are small birds of Spain and Morocco. Males of the species demonstrate an exaggerated sexual display by carrying many heavy stones to nesting cavities. This 35-gram bird transports, on average, 3.1 kg of stones per nesting season! Different males carry somewhat different sized stones, prompting a study of whether larger stones may be a signal of higher health status. M. Soler et al. (“Weight Lifting and Health Status in the Black Wheatear,” *Behavioral Ecology* 10(3) (1999): 281–86) calculated the average stone mass (g) carried by each of 21 male black wheatears, along with T-cell response measurements reflecting their immune systems’ strengths. The data in Display 7.19 were taken from their Figure 1. Analyze the data and write a statistical report summarizing the evidence supporting whether health—as measured by T-cell response—is associated with stone mass, and quantifying the association.

DISPLAY 7.19

Mass of stones carried and immune system strength for 21 wheatear birds; first 5 of 15 rows

Bird	Mean stone mass (g)	T-cell response (mm)
1	3.33	0.252
2	4.62	0.263
3	5.43	0.251
4	5.73	0.251
5	6.12	0.183

28. Brain Activity in Violin and String Players. Studies over the past two decades have shown that activity can effect the reorganization of the human central nervous system. For example, it is known that the part of the brain associated with activity of a finger or limb is taken over for other purposes in individuals whose limb or finger has been lost. In one study, psychologists used magnetic source imaging (MSI) to measure neuronal activity in the brains of nine string players (six violinists, two cellists, and one guitarist) and six controls who had never played a musical instrument, when the thumb and fifth finger of the left hand were exposed to mild stimulation. The researchers felt that stringed instrument players, who use the fingers of their left hand extensively, might show different behavior in the brain—as a result of this extensive physical activity—than individuals who did not play stringed instruments. Display 7.20 shows a neuron activity index from the MSI and the years that the individual had been playing a stringed instrument (zero for the controls). (Data based on a graph in Elbert et al., “Increased Cortical Representation of the Fingers of the Left Hand in String Players,” *Science* 270 (13 October, 1995) 305–7.) Is the neuron activity different in the stringed musicians and the controls? Is the amount of activity associated with the number of years the individual has been playing the instrument?

29. Sampling Bias in Exit Polls. Exit pollsters predict election results before final counts are tallied by sampling voters leaving voting locations. The pollsters have no way of selecting a random sample, so they instruct their interviewers to select every third exiting voter, or fourth, or tenth, or some other specified number. Some voters refuse to participate or avoid the interviewer. If the refusers and avoiders have the same voting patterns as the rest of the population, then it shouldn’t matter; the sample, although not random, wouldn’t be biased. If, however, one candidate’s voters are more likely to refuse or avoid interview, the sample would be biased and could lead to a misleading conclusion.

DISPLAY 7.20

Years that the individual has been playing a stringed instrument and neuronal activity index ("D5 dipole strength, in nA·m") for nine stringed musicians and six controls

Years playing	Neuron activity index
0	5
0	6
0	7.5
0	9
0	9.5
0	11
5	16
6	16.5
8	11.5
10	16
12	25
13	25.5
17	25.5
18	23
19	26.5

On November 4, 2004, exit pollsters incorrectly predicted that John Kerry would win the U.S. presidential election over George W. Bush. The exit polls overstated the Kerry advantage in 42 of 50 states. No one expects exit polls to be exact, but chance alone cannot reasonably explain this discrepancy. Although fraud is a possibility, the data are also consistent, with Bush supporters being more likely than Kerry supporters to refuse or avoid participation in the exit poll.

In a postelection evaluation, the exit polling agency investigated voter avoidance of interviewers. Display 7.21 shows the average Kerry exit poll overestimate (determined after the actual counts were available) for a large number of voting precincts, grouped according to the distance of the interviewer from the door. If Bush voters were more likely to avoid interviewers in general, one might also expect a greater avoidance with increasing distance to the interviewer (since there is more opportunity for escape). A positive relationship between distance of the interviewer from the door and amount of Kerry overestimate, therefore, would lend credibility to the theory that Bush voters were more likely to avoid exit poll interviews. How strong is the evidence that the mean Kerry overestimate increases with increasing distance of interviewer from

DISPLAY 7.21

Exit poll error in favor of Kerry and distance of exit poll interviewer from the voting precinct door, in the 2004 U.S. presidential election between George W. Bush and John Kerry

Overestimate	Distance
5.3	0
6.4	5
5.6	17
7.6	37
9.6	75
12.3	100

DISPLAY 7.22

Average exit poll interview refusal rates for precincts grouped according to the approximate age of the interviewer in the 2004 U.S. presidential election between George W. Bush and John Kerry

Age	Refusal
22	0.39
30	0.38
40	0.35
50	0.32
60	0.31
65	0.29

the door? (Data from Evaluation of Edison/Mitofsky Election System 2004 prepared by Edison Media Research and Mitofsky International for the National Election Pool (NEP), January 15, 2005. <http://abcnews.go.com/images/Politics/EvaluationofEdisonMitofskyElectionSystem.pdf> (accessed May 9, 2008).)

30. Sampling Bias in Exit Polls 2. This exercise is about differential interview *refusal* rates in the exit polls conducted in the 2004 U.S. presidential election. Display 7.22 shows the average proportion of voters who refused to be interviewed at precincts grouped according to the approximate age of the interviewer. What evidence do these data provide that the mean refusal rate decreased with increasing age of interviewer? An affirmative answer to this question doesn't provide any direct evidence of a difference between Kerry and Bush voters, but is consistent with an undercount of Bush votes in the exit polls if, as one might speculate based on the relative conservativeness of Bush supporters, Bush voters were more likely to avoid younger interviewers. (See Exercise 29.)

Answers to Conceptual Exercises

1. (a) Yes. (b) Yes. (c) Such an extrapolation would be risky.
2. The standard deviation σ about the regression reflects measurement error variation. Making this smaller will cause the standard deviations of the sampling distributions of the least squares estimates to be smaller (see Display 7.7).
3. The model says that the distribution is normal. The estimated mean pH is about 6.6. The prediction limits will be about 2 SDs up and down, so the SD is about 0.1 pH units.
4. This implies an exact relationship between Y and X . The model should be for the *mean* of Y as a function of X .
5. (a) kg; (b) kg/cm; (c) kg; (d) none.
6. Regression refers to the mean of a response variable as a function of an explanatory variable. A regression model is a function used to describe the regression. The simple linear regression model is a particular regression model in which the regression is a straight-line function of a single explanatory variable.
7. None.
8. This is the regression effect. It is exactly what you can expect to happen.
9. (a) No. Height = 0 is outside the range of observed values, so the model may not extend to that situation. (b) No. It may be useful for answering questions pertaining to the regression of weight on height for heights in a certain range.
10. (a) At the sample average of the X 's used in the estimation. (b) Same as (a).
11. σ .

A Closer Look at Assumptions for Simple Linear Regression

The inferential tools of the previous chapter are based on the normal simple linear regression model with constant variance. Since real data do not necessarily conform to this model, the data analyst must size up the situation and choose a course of action based on an understanding of the robustness of the tools to model violations.

This chapter presents some informal graphical tools and a formal test for assessing the lack of fit. As before, the graphical procedures are used to find suitable transformations to scales where the simple linear regression model seems appropriate. The lack-of-fit test, on the other hand, looks specifically at the issue of whether the straight line assumption is plausible.

8.1 CASE STUDIES

8.1.1 Island Area and Number of Species—An Observational Study

Biologists have noticed a consistent relation between the area of islands and the number of animal and plant species living on them. If S is the number of species and A is the area, then $S = CA^\gamma$ (roughly), where C is a constant and γ is a biologically meaningful parameter that depends on the group of organisms (birds, reptiles, or grasses, for example). Estimates of this relationship are useful in conservation biology for predicting species extinction rates due to diminishing habitat.

The data in Display 8.1 are the numbers of reptile and amphibian species and the island areas for seven islands in the West Indies. (Data on species from E. O. Wilson, *The Diversity of Life*, New York: W. W. Norton, 1992; areas from *The 1994 World Almanac*, Mahwah, N.J.: Funk & Wagnalls, 1993.) These are typical of data used to estimate the area effect. Researchers wish to estimate γ in the species-area equation for this group of animals and to summarize the effect of area on the typical number of species.

DISPLAY 8.1 Island area and number of reptile and amphibian species for seven islands in the West Indies

Island	Area (square miles)	Number of species
Cuba	44,218	100
Hispaniola	29,371	108
Jamaica	4,244	45
Puerto Rico	3,435	53
Montserrat	32	16
Saba	5	11
Redonda	1	7

Statistical Conclusion

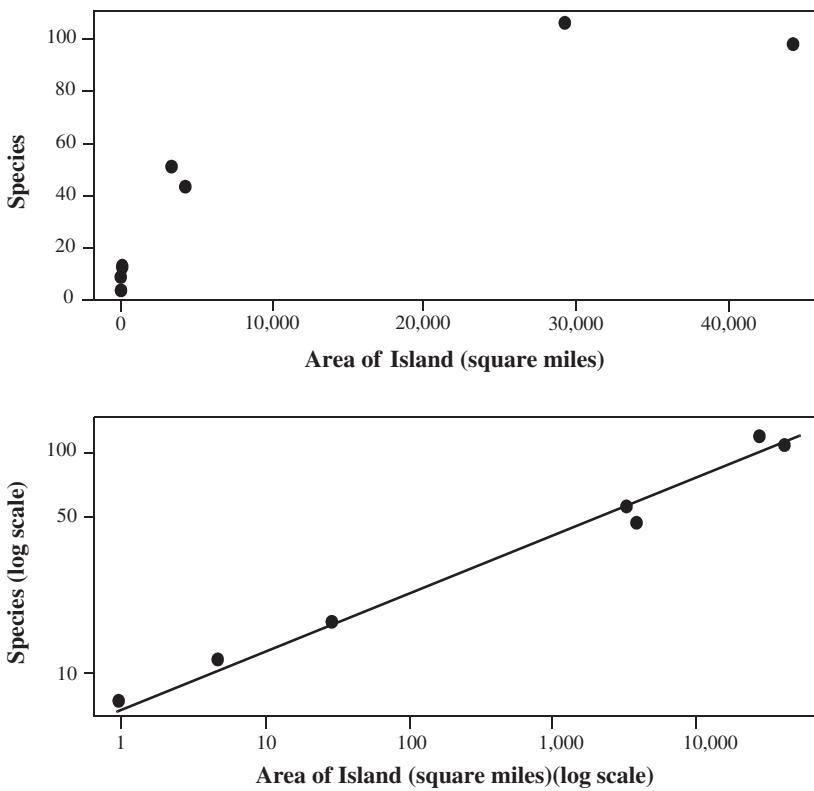
A second graph in Display 8.2 is a log-log scatterplot of number of species versus island area, along with the estimated line for the regression of log number of species on log area. The parameter γ in the species-area relation, Median $\{S|A\} = CA^\gamma$, is estimated to be 0.250 (a 95% confidence interval is 0.219 to 0.281). It is estimated that the median number of species increases by 19% with each doubling of area.

Scope of Inference

The statistical association from these observational data cannot be used to establish a causal connection. Furthermore, any generalization of these results to islands in other parts of the world or to a wider population, like “islands of rain forest,” is purely speculative. Nevertheless, the results from these data are consistent with results for other islands and with small-scale randomized experiments. In

DISPLAY 8.2

Scatterplot and log-log-scatterplot of number of reptile and amphibian species versus area for seven islands in the West Indies



summarizing the studies, Wilson offered a conservatively optimistic guess that the current rate of environmental destruction amounts to 27,000 species vanishing each year and that 20% of all plant and animal species currently on earth will be extinct by the year 2022.

8.1.2 Breakdown Times for Insulating Fluid Under Different Voltages—A Controlled Experiment

In an industrial laboratory, under uniform conditions, batches of electrical insulating fluid were subjected to constant voltages until the insulating property of the fluids broke down. Seven different voltage levels, spaced 2 kilovolts (kV) apart from 26 to 38 kV, were studied. The measured responses were the times, in minutes, until breakdown, as listed in Display 8.3. (Data from W. B. Nelson, Schenectady, N.Y.: GE Co. Technical Report 71-C-011 (1970), as discussed in J. F. Lawless, *Statistical Models and Methods for Lifetime Data*, New York: John Wiley & Sons, 1982, chap. 4.) How does the distribution of breakdown time depend on voltage?

DISPLAY 8.3

Times (in minutes) to breakdown of 76 samples of an insulating fluid subjected to different constant voltages

Group #:	1	2	3	4	5	6	7
Voltage (kV):	26	28	30	32	34	36	38
Sample size:	3	5	11	15	19	15	8
Times (min):	5.79 1579.52 2323.70	68.85 108.29 110.29 426.07 1067.60	7.74 17.05 20.46 21.02 22.66 3.91 9.88 13.95 15.93 27.80 53.24 82.85 89.29 100.59 215.10	0.27 0.40 0.69 0.79 2.75 3.16 4.15 4.67 4.85 6.50 7.35 8.01 8.27 12.06 31.75	0.19 0.78 0.96 1.31 2.78 1.99 2.07 2.58 2.71 2.90 3.67 3.99 5.35 13.77 25.50	0.35 0.59 0.96 0.99 1.69 1.97 2.07 2.38	0.09 0.39 0.47 0.73 0.74 1.13 1.40 2.38

Statistical Conclusion

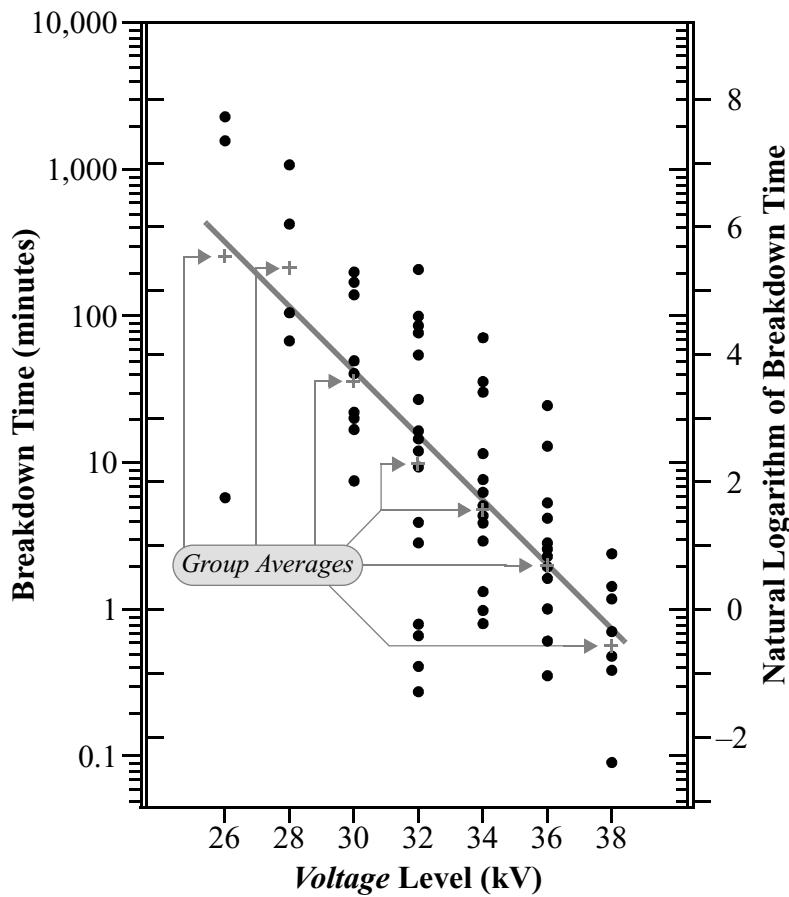
Display 8.4 shows a scatterplot of the responses on a logarithmic scale plotted against the voltage levels, along with a summary of the parameter estimates in the simple linear regression of log breakdown time on voltage. The median breakdown time decreases by 40% for each 1 kV increase in voltage (95% confidence interval: 32% to 46%).

Scope of Inference

The laboratory setting for this experiment allowed the experimenter to hold all factors constant except the voltage level, which was assigned at different levels to different batches. Therefore it seems reasonable to infer that the different voltage levels must be directly responsible for the observed differences in time to breakdown. It can be inferred that other batches in the same laboratory setting would follow the same pattern so long as the voltage level is within the experimental range. Inference to voltage levels outside the experimental range and to performance under nonlaboratory conditions cannot be made from these data. If such inference is required, further testing or stronger assumptions must be invoked.

DISPLAY 8.4

Scatterplot of breakdown times (natural logarithm scale) versus voltage levels and summary of estimated simple linear regression



8.2 ROBUSTNESS OF LEAST SQUARES INFERENCES

Exact justification of the statistical statements—tests, confidence intervals, and prediction intervals—based on least squares estimates depends on these features of the regression model:

1. **Linearity.** The plot of response means against the explanatory variable is a straight line.

2. **Constant variance.** The spread of the responses around the straight line is the same at all levels of the explanatory variable.
3. **Normality.** The subpopulations of responses at the different values of the explanatory variable all have normal distributions.
4. **Independence.** The location of any response in relation to its mean cannot be predicted, either fully or partially, from knowledge of where other responses are in relation to their means. (Furthermore, the location of any response in relation to its mean cannot be predicted from knowledge of the explanatory variable values.)

The Linearity Assumption

Two violations of the first assumption may occur: A straight line may be an inadequate model for the regression (the regression might contain some curvature, for example); or a straight line may be appropriate for most of the data, but contamination from one or several outliers from different populations may render it inapplicable to the entire set. Both of these violations can cause the least squares estimates to give misleading answers to the questions of interest. Estimated means and predictions can be biased—they systematically under- or overestimate the intended quantity—and tests and confidence intervals may inaccurately reflect uncertainty. Although the severity of the consequences is always related to the severity of the violation, it is undesirable to use simple linear regression when the linearity assumption is not met. Remedies are available for dealing with this situation (see Chapter 9).

The Equal-Spread Assumption

The consequences for violating this assumption are the same as for one-way analysis of variance. Although the least squares estimates are still unbiased even if the variance is nonconstant, the standard errors inaccurately describe the uncertainty in the estimates. Tests and confidence intervals can be misleading.

The Normality Assumption

Estimates of the coefficients and their standard errors are robust to nonnormal distributions. Although the tests and confidence intervals originate from normal distributions, the consequences of violating this assumption are usually minor. The only situation of substantial concern is when the distributions have long tails (outliers are present) and sample sizes are moderate to small.

If prediction intervals are used, on the other hand, departures from normality become important. This is because the prediction intervals are based directly on the normality of the *population distributions* whereas tests and confidence intervals are based on the normality of the *sampling distributions of the estimates* (which may be approximately normal even when the population distributions are not).

The Independence Assumption

Lack of independence causes no bias in least squares estimates of the coefficients, but standard errors are seriously affected. As before, cluster and serial effects, if

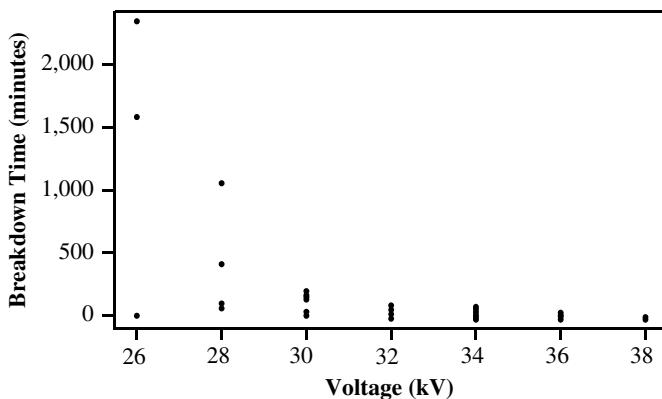
suspected, should be incorporated with more sophisticated models. See Chapters 13 and 14 for incorporating cluster effects using multiple regression and Chapter 15 for ways to incorporate serial effects into regression models.

8.3 GRAPHICAL TOOLS FOR MODEL ASSESSMENT

The principal tools for model assessment are scatterplots of the response variable versus the explanatory variable and of the residuals versus the fitted values. An initial scatterplot of species number versus island area, on the top of Display 8.2, immediately indicates problems with a straight line model for the means. Similarly, a scatterplot of breakdown time versus voltage, in Display 8.5, shows problems of nonlinearity and nonconstant variance.

DISPLAY 8.5

Scatterplot of breakdown times versus voltage



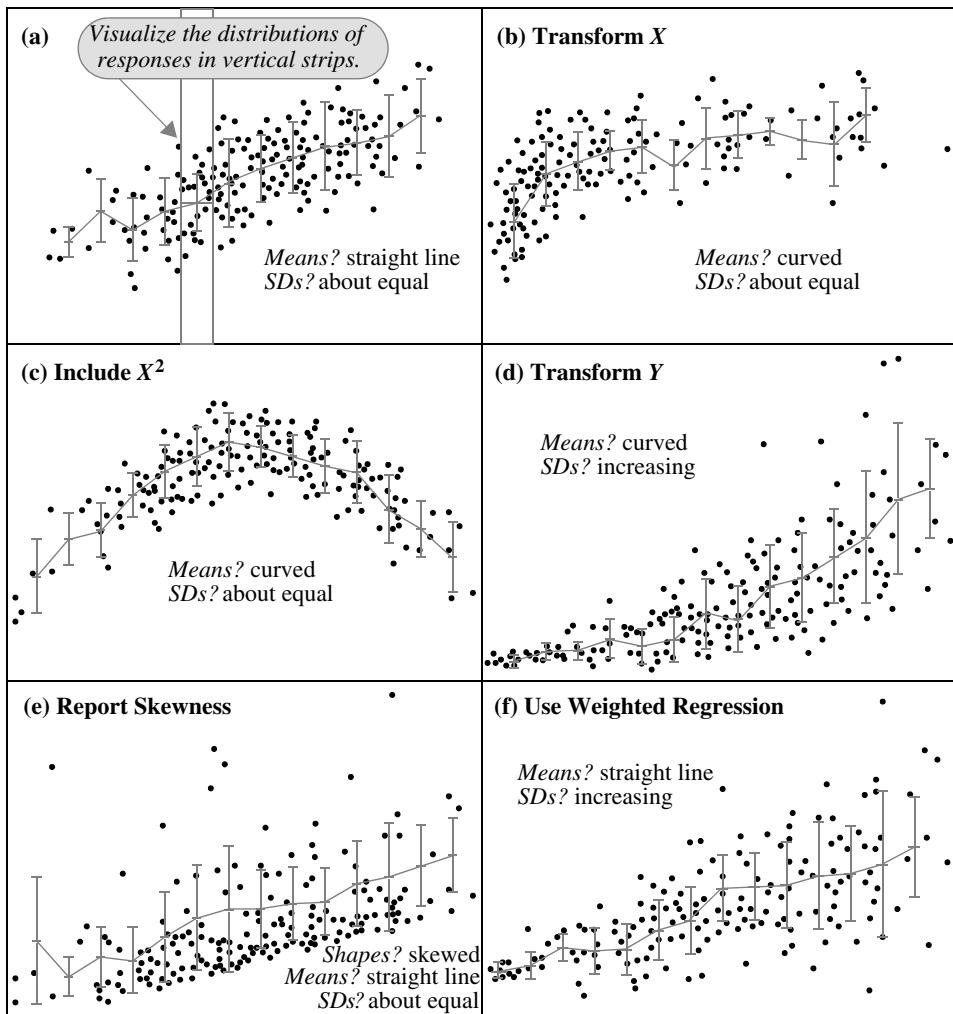
8.3.1 Scatterplot of the Response Variable Versus the Explanatory Variable

Scatterplots with some commonly occurring patterns are shown in Display 8.6. The eye should group points into vertical strips, as shown in (a), to explore the characteristics of subpopulation distributions. Lines connecting strip averages roughly exhibit the relationship between the mean of the responses and the explanatory variable. Vertical lines show the strip standard deviations, which should be examined for patterns in variability. As discussed for each of the plots, the particular patterns for the mean and the variability often suggest a course of action.

- (a) This is the ideal situation. The regression is a straight line and the variability is about the same at all locations along the line. Use simple linear regression.
- (b) The regression is not a straight line, but it is monotonic (the mean of Y either strictly increases or strictly decreases in X), and the variability is about the

DISPLAY 8.6

Some hypothetical scatterplots of response versus explanatory variable with suggested courses of action; (a) is ideal



same at all values of X . Try transforming X to a new scale where the straight line assumption appears defensible. Then use simple linear regression.

- (c) The regression is not a straight line and is not monotonic, and the variability is about the same at all values of X . No transformation of X can yield a straight line relationship. Try *quadratic regression* ($\mu\{Y|X\} = \beta_0 + \beta_1 X + \beta_2 X^2$), which is discussed in Chapter 9.
- (d) The regression is not a straight line, and the variability increases as the mean of Y increases. Try a transformation of Y , such as log, reciprocal, or square root, followed by simple linear regression.

- (e) The regression is a straight line, the variability is roughly constant, but the distribution of Y about the regression line is skewed. Remedies are unnecessary, and transformations will create other problems. Use simple linear regression, but report the skewness.
- (f) The regression is a straight line but the variability increases as the mean of Y increases. Simple linear regression gives unbiased estimates of the straight line relationship, but better estimates are available using *weighted regression*, as discussed in Section 11.6.1.

8.3.2 Scatterplots of Residuals Versus Fitted Values

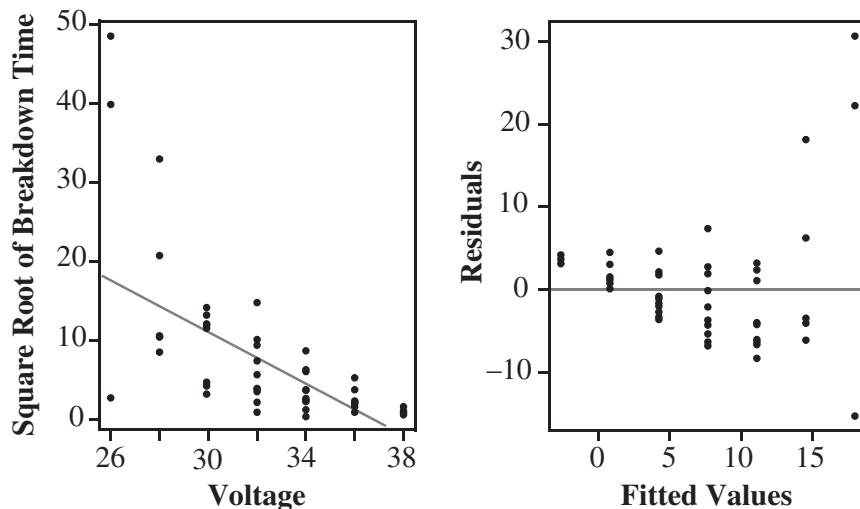
Sometimes the patterns in Display 8.6 are difficult to detect because the total variability of the response variable is much larger than the variability around the regression line. Scatterplots of the residuals versus the fitted values are better for finding patterns because the linear component of variation in the responses has been removed, leaving a clearer picture about curvature and spread. The residual plot alerts the user to *nonlinearity, nonconstant variance, and the presence of outliers*.

Example

Display 8.5 is a scatterplot of the breakdown time versus voltage. This scatterplot resembles the pattern in part (d) of Display 8.6, suggesting the need for transformation. Display 8.7 shows the scatterplot and residual plot after applying a square root transformation to the response. The residual plot is more informative than the scatterplot. It has a classic *horn-shaped* pattern, showing a combination of a

DISPLAY 8.7

Scatterplot of the square root of breakdown time versus voltage and a residual plot based on the simple linear regression fit



poor fit to the subpopulation averages and increasing variability. It is evident that the square root transformation has not resolved the problem. A log transformation, however, works well (see Display 8.4). (*Note:* The horn-shaped pattern is always a key indicator of the need to transform, whether or not there are replicate samples at specific values of the explanatory variable.)

Transformations Indicated by Horn-Shaped Residual Plots

A horn-shaped pattern in the residual plot suggests a response transformation like the square root, the logarithm, or the reciprocal. The logarithm is the easiest to interpret. The reciprocal, $1/Y$, works better when the nonconstant spread is more severe, and the square root works better when the nonconstant spread is less severe. Judging severity of nonconstant variance in the horn-shaped residual plot is difficult and unnecessary, however. Try one of the transformations, re-fit the regression model on the transformed data, and redraw the residual plot to see if the transformation has worked. If not, then one of the others can be attempted.

8.4 INTERPRETATION AFTER LOG TRANSFORMATIONS

A data analyst must interpret regression results in a way that makes sense to the intended audience. The appropriate wording for inferential statements after logarithmic transformation depends on whether the transformation was applied to the response, to the explanatory variable, or to both.

When the Response Variable Is Logged

If $\mu\{\log(Y)|X\} = \beta_0 + \beta_1 X$, and if the distribution of the transformed responses about the regression is symmetric, then

$$\text{Median}\{Y|X\} = \exp(\beta_0)\exp(\beta_1 X).$$

Consequently,

$$\text{Median}\{Y|(X + 1)\}/\text{Median}\{Y|X\} = \exp(\beta_1),$$

so an increase in X of 1 unit is associated with a multiplicative change of $\exp(\beta_1)$ in $\text{Median}\{Y|X\}$.

For example, the estimated relationship between the breakdown time (BDT) of insulating fluid and voltage is $\hat{\mu}\{\log(\text{BDT}) | \text{Voltage}\} = 19.0 - 0.51 \text{ Voltage}$. A 1 kV increase in voltage is associated with a multiplicative change in median BDT of $\exp(-0.51)$, or 0.60. So, the median breakdown time at 28 kV is 60% of what it is at 27 kV; the median breakdown time at 29 kV is 60% of what it is at 28 kV, and so on. Since a 95% confidence interval for β_1 is -0.62 to -0.39 , a 95% confidence interval for $\exp(\beta_1)$ is $\exp(-0.62)$ to $\exp(-0.39)$, or 0.54 to 0.68.

The statement $\text{Median}\{Y|(X + 1)\} = 0.6 \times \text{Median}\{Y|X\}$ can also be written as $\text{Median}\{Y|(X + 1)\} - \text{Median}\{Y|X\} = 0.4 \times \text{Median}\{Y|X\}$, which permits the following kind of statement: "It is estimated that the median of Y decreases by 40% for each one unit increase in X (95% confidence interval: 32% to 46%)."

When the Explanatory Variable Is Logged

The relationship $\mu\{Y|\log(X)\} = \beta_0 + \beta_1 \log(X)$ can be described in terms of multiplicative changes in X , either as a change in the mean of Y for each doubling of X or a change in the mean of Y for each ten-fold increase in X . The chosen multiple should be consistent with the range of X 's in the data set.

Notice that

$$\mu\{Y|\log(2X)\} - \mu\{Y|\log(X)\} = \beta_1 \log(2),$$

so a doubling of X is associated with a $\beta_1 \log(2)$ change in the mean of Y . Similarly, a ten-fold increase in X is associated with a $\beta_1 \log(10)$ change in the mean of Y .

For the meat processing data of Section 7.1.2, $\hat{\mu}\{\text{pH}|\log(\text{Time})\} = 6.98 - 0.726 \log(\text{Time})$, so a doubling of time after slaughter is associated with a $\log(2)(-0.726) = -0.503$ unit change in pH. Since a 95% confidence interval for β_1 is from -0.805 to -0.646 , a 95% CI for $\log(2)\beta_1$ is from -0.558 to -0.448 . In words: It is estimated that the mean pH is reduced by 0.503 for each doubling of time after slaughter (95% confidence interval 0.448 to 0.558).

When Both the Response and Explanatory Variables Are Logged

The interpretation is a combination of the previous two. If $\mu\{\log(Y)|\log(X)\} = \beta_0 + \beta_1 \log(X)$, then $\text{Median}\{Y|X\} = \exp(\beta_0)X^{\beta_1}$. A doubling of X is associated with a multiplicative change of 2^{β_1} in the median of Y . Or, a ten-fold increase in X is associated with a 10^{β_1} -fold change in the median of Y .

For the island size and number of species data, $\hat{\mu}\{\log(\text{species})|\log(\text{area})\} = 1.94 + 0.250 \log(\text{area})$. Thus, an island area of $2A$ is estimated to have a median number of species that is $2^{0.250}$ (or 1.19) times the median number of species for an island of area A . Since a 95% confidence interval for β_1 is 0.219 to 0.281, a 95% confidence interval for the multiplicative factor in the median is $2^{0.219}$ to $2^{0.281}$, or 1.16 to 1.22.

The Need for Interpretation

An important aspect of the log transformation is that straightforward multiplicative statements emerge. Straightforward interpretations after other transformations only follow in certain instances. For example, the square root of the cross-sectional area of a tree may be re-expressed as the diameter; the reciprocal of the time to complete a race may be interpreted as the speed. In general, however, interpretation for other transformations may be awkward.

For two types of questions, interpretation is not critical. First, if the regression is used only to assess whether the distribution of the response is *associated with* the explanatory variable, it is sufficient to test the hypothesis that the slope in the regression of Y on X is zero, where Y or X are transformed variables. If the

distribution of the transformed response is associated with X , the distribution of the response is associated with X , even though that association may be difficult to describe. Secondly, if the purpose is prediction, no interpretation of the regression coefficients is needed. It is only necessary to express the prediction on the original scale, regardless of the expression used to make the prediction.

8.5 ASSESSMENT OF FIT USING THE ANALYSIS OF VARIANCE

An analysis of variance can be used to compare several models. When there are replicate response variables at several explanatory variable values, an analysis of variance F -test for comparing the simple linear regression model to the separate-means (one-way analysis of variance) model supplies a formal assessment of the goodness of fit of simple linear regression. This is called the *lack-of-fit F-test*.

8.5.1 Three Models for the Population Means

In the following three model descriptions for means, the subscript i refers to the group number (e.g., different voltage levels).

1. Separate-means model: $\mu\{Y|X_i\} = \mu_i$, for $i = 1, \dots, I$.
2. Simple linear regression model: $\mu\{Y|X_i\} = \beta_0 + \beta_1 X_i$, for $i = 1, \dots, I$.
3. Equal-means model: $\mu\{Y|X_i\} = \mu$, for $i = 1, \dots, I$.

The separate-means model has no restriction on the values of any of the means. It has I different parameters—the individual group means. The simple linear regression model has two parameters, the slope and the intercept. The equal-means model has a single parameter.

These models form a *hierarchical* set. The equal-means model is a special case of the simple linear regression model, which in turn is a special case of the separate-means model. Viewed conversely, the separate-means model is a generalization of the simple linear regression model, which in turn is a generalization of the equal-means model. A generalization of one model is another model that contains the same features but uses additional parameters to describe a more complex structure.

8.5.2 The Analysis of Variance Table Associated with Simple Regression

Display 8.8 contains two different analysis of variance tables. Table (b) comes from a one-way analysis of variance (Chapter 5), showing the details of an F -test that compares the separate-means model to the equal-means model for the insulating fluid example of Section 8.1.2. Table (a) comes from a simple linear regression analysis showing the details of an F -test that compares the simple linear regression model to the equal-means model.

If β_1 is zero, the simple linear regression model reduces to the equal-means model, $\mu\{Y|X\} = \beta_0$. The hypothesis that $\beta_1 = 0$ can therefore be tested with a comparison of the sizes of the residuals from fits to these two models. An analysis of

DISPLAY 8.8

Analysis of variances tables for the insulating fluid data from a simple linear regression analysis and from a separate-means (one-way ANOVA) analysis

(a): Analysis of variance table from a simple linear regression analysis

Source	Sum of squares	d.f.	Mean square	F-statistic	p-value
Regression	190.1514	1	190.1514	78.14	<0.0001
Residual	180.0745	74	2.4334		
Total	370.2258	75			

Annotations below the table:

- Residual sum of squares, regression model* (points to the Residual row)
- $\hat{\sigma}^2$ in regression model* (points to the Mean square column)
- compares regression and equal-means models* (points to the p-value column)

(b): Analysis of variance table from a one-way analysis of variance

Source	Sum of squares	d.f.	Mean square	F-statistic	p-value
Between groups	196.4774	6	32.7462	13.00	<0.0001
Within groups	173.7484	69	2.5181		
Total	370.2258	75			

Annotations below the table:

- Residual sum of squares, separate-means model* (points to the Within groups row)
- $\hat{\sigma}^2$ in separate-means model* (points to the Mean square column)
- compares separate-means and equal-means models* (points to the p-value column)

variance table associated with simple linear regression displays the components of the F -statistic for this test (Display 8.8(a)). Its two main features are the following: The p -value is the same as the two-sided p -value for the t -test of $H_0: \beta_1 = 0$ (the F -statistic is the square of the t -statistic); and the residual mean square is the estimate of variance, $\hat{\sigma}^2$.

The sums of squares column contains the important working pieces. In Table (a), the *residual sum of squares* is the sum of squares of residuals from the regression model, while the *total sum of squares* is the sum of squares of residuals from the equal-means model. The *regression sum of squares* is their difference; that is, it is the amount by which the residual sum of squares decreases when the model for the mean of Y is generalized by adding $\beta_1 X$. The generalization adds one parameter to the model, and that is the number of degrees of freedom associated with the regression sum of squares.

Except for differences in terminology, this is completely analogous to the sums of squares column in Table (b), where the *between sum of squares* is the amount by which the residual sum of squares is reduced by the generalization from the

equal-means to the separate-means model, involving the change from one to seven parameters (difference = 6 parameters added).

8.5.3 The Lack-of-Fit F -Test

When replicate values of the response occur at some or all of the explanatory variable values, a formal test of the adequacy of the straight-line regression model is available. The test is an extra-sum-of-squares F -test comparing the simple linear (reduced) model to the separate means (full) model. (See Sections 5.3 and 5.4.)

Specifically, the lack-of-fit F -statistic is

$$F_{\text{stat}} = \frac{[\text{SSRes}_{\text{LR}} - \text{SSRes}_{\text{SM}}]/[\text{d.f.}_{\text{LR}} - \text{d.f.}_{\text{SM}}]}{\hat{\sigma}_{\text{SM}}^2},$$

where SSRes_{LR} and SSRes_{SM} are the sums of squares of residuals from the simple linear regression (“Residual” in Display 8.8(a)) and separate-means models (“Within groups” in Display 8.8(b)), respectively; and d.f._{LR} and d.f._{SM} are the degrees of freedom associated with these residual sums of squares. The denominator of the F -statistic is the estimate of σ^2 from the separate-means model. The p -value, for a test of the null hypothesis that the simple linear regression model fits, is found as the proportion of values from an F distribution that exceed the F -statistic. The numerator degrees of freedom are $\text{d.f.}_{\text{LR}} - \text{d.f.}_{\text{SM}}$, and the denominator degrees of freedom are d.f._{SM} .

Test Computation

Some statistical computer packages will automatically compute the lack-of-fit F -test statistic within the simple regression procedure, as long as there are indeed replicates. Others might perform this test if requested. If neither of these options is available, the user must obtain the analysis of variance tables from both the regression fit and the one-way analysis of variance, as in Display 8.8, extract the necessary information, and compute the F -statistic manually.

Example—Insulating Fluid

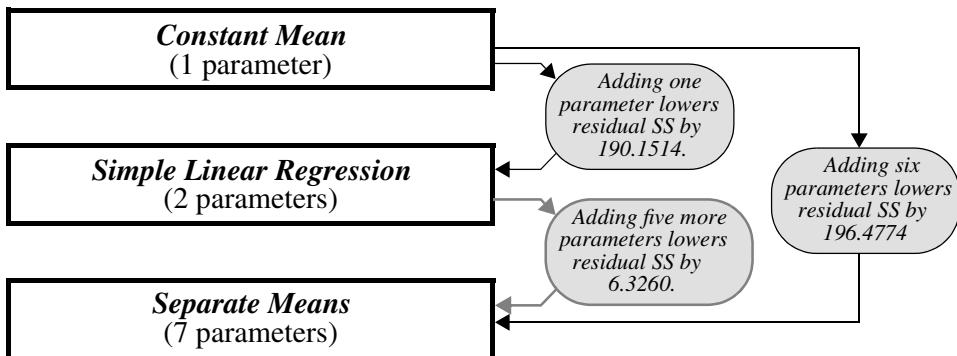
From Display 8.8, $\text{SSRes}_{\text{LR}} = 180.0745$, $\text{SSRes}_{\text{SM}} = 173.7484$, $\text{d.f.}_{\text{LR}} = 74$, and $\text{d.f.}_{\text{SM}} = 69$. So the F -statistic is $[(180.0745 - 173.7484)/(74 - 69)]/2.5181 = 0.502$. The proportion of values from an F -distribution on 5 and 69 degrees of freedom that exceed 0.502 is 0.78. This large p -value provides no evidence of lack-of-fit to the simple linear regression model. A small p -value would have suggested that the variability between group means cannot be explained by a simple linear regression model.

8.5.4 A Composite Analysis of Variance Table

Since the simple linear regression model is intermediate between the equal-means and the separate-means models, the reduction in residual sum of squares (196.4774) associated with generalizing from the equal-means model to the separate-means

DISPLAY 8.9

Reductions in sums of squared residuals in hierarchical models for mean responses in the insulating fluid study

**DISPLAY 8.10**

Composite analysis of variance table with *F*-test for lack of fit

Source of variation	Sum of squares	d.f.	Mean square	<i>F</i> -statistic	p-value
Between groups	196.4774	6	32.7462	13.00	<0.0001
Regression	190.1514	1	190.1514	75.51	<0.0001
Lack of fit	6.3260	5	1.2652	0.50	0.78
Within groups	173.7484	69	2.5181		
Total	370.2258	75			

LEGEND

Normal type items come from regression analysis (a).
Italicized items come from separate-means analysis (b).
Boldface items are new and calculated here.

By subtraction

model can be broken up into contributions associated with an initial generalization to the simple linear regression model (190.1514) and with a further generalization from there to the separate-means model. (See Display 8.9.) The amount associated with the latter step is the difference between the two, $196.4774 - 190.1514 = 6.3260$. Therefore, a useful composite of the analysis of variance tables (a) and (b) in Display 8.8 is given in Display 8.10.

Essentially this is the same table as the analysis of variance table for the separate-means model, Display 8.8 (b), except that the “Between group” sum of squares has been broken into two components—one that measures the pattern in the means that follows a straight line, and one that measures any patterns that fail to follow the straight line model. The latter is called the *lack-of-fit* component. The mean squares in each row are the sum of squares divided by degrees of freedom.

8.6 RELATED ISSUES

8.6.1 R-Squared: The Proportion of Variation Explained

The *R-squared* statistic, or *coefficient of determination*, is the percentage of the total response variation explained by the explanatory variable. Referring to the analysis of variance table (a) in Display 8.8, the residual sum of squares for the equal-means model, $\mu\{Y\} = \beta_0$, is 370.2258. This measures total response variation. The residual sum of squares for the simple linear regression model $\mu\{Y|X\} = \beta_0 + \beta_1 X$ is 180.0745. This residual sum of squares measures the response variation that remains unexplained after inclusion of the $\beta_1 X$ term in the model. Including the explanatory variable therefore reduced the variability by 190.1513. *R-squared* expresses this reduction as a percentage of total variation:

$$R^2 = 100 \left(\frac{\text{Total sum of squares} - \text{Residual sum of squares}}{\text{Total sum of squares}} \right) \%$$

For the insulating fluids example,

$$R^2 = 100 \left(\frac{370.2258 - 180.0745}{370.2258} \right) \% = 51.4\%.$$

This should be read as “Fifty-one percent of the variation in log breakdown times was explained by the linear regression on voltage.” The statement is phrased in the past tense because R^2 describes what happened in the analyzed data set.

Notice that if the residuals are all zero (a perfect fit), then R^2 is 100%. At the other extreme, if the best-fitting regression line has slope 0 (and therefore intercept \bar{Y}), the residuals will be exactly equal to $Y_i - \bar{Y}$, the residual sum of squares will be exactly equal to the total sum of squares, and R^2 will be zero.

A judgment about what constitutes “good” values for R^2 depends on the context of the study. In precise laboratory work, R^2 values under 90% may be low enough to require refinements in technique or the inclusion of other explanatory information. In some social science contexts, however, where a single variable rarely explains a great deal of the variation in a response, R^2 values of 50% may be considered remarkably good.

For simple linear regression, R^2 is identical to the square of the sample correlation coefficient for the response and the explanatory variable. As with the correlation, R^2 only estimates some population quantity if the (X, Y) pairs are randomly drawn from a population of pairs. It should not be used for inference, and it should never be used to assess the adequacy of the straight line model, because R^2 can be quite large even when the simple linear regression model is inadequate.

DISPLAY 8.11

Estimates of group means and standard errors using different approaches

Voltage Level (kV)	n	Average log (BDT)	Internal SE	Pooled SE	Regression	
					Estimate	SE
26	3	5.6240	1.9371	0.9162	5.7640	0.4467
28	5	5.3295	0.5119	0.7907	4.7492	0.3446
30	11	3.8220	0.3350	0.4785	3.7345	0.2536
32	15	2.2285	0.5675	0.4097	2.7198	0.1904
34	19	1.7864	0.3499	0.3640	1.7050	0.1858
36	15	0.9022	0.2866	0.4097	0.6903	0.2432
38	8	-0.4243	0.3506	0.5610	-0.3244	0.3318

Means estimated by:
 (1) sample averages
 (2) regression model

Regression

Accuracy estimated by:
 (i) sample SD/ \sqrt{n}
 (ii) pooled SD/ \sqrt{n}
 (iii) regression method

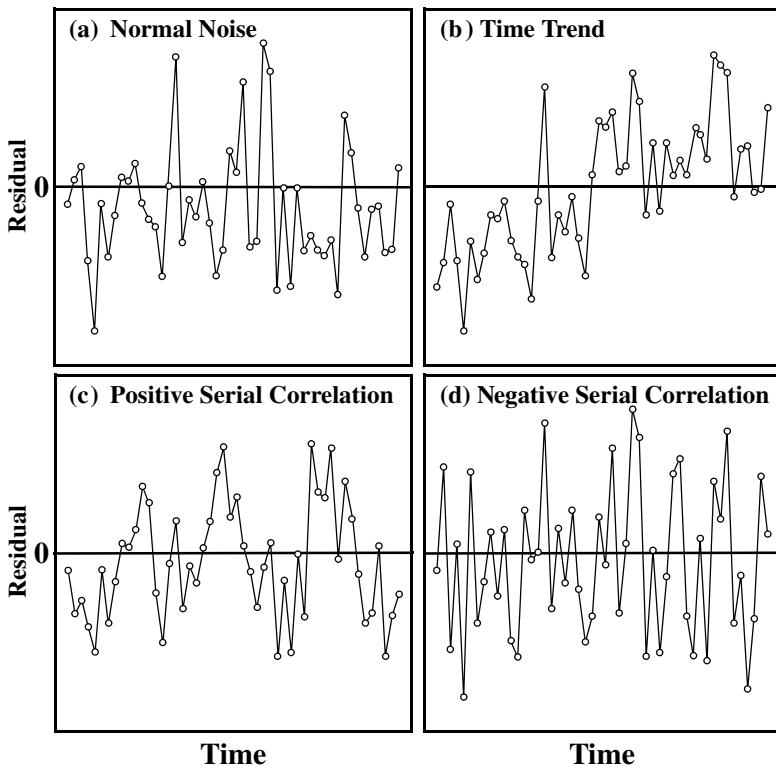
8.6.2 Simple Linear Regression or One-Way Analysis of Variance?

When the data are arranged in groups that correspond to different levels of an explanatory variable (as in the meat processing and insulating fluid studies), the statistical analysis may be based on either simple linear regression or one-way analysis of variance. The choice between these two techniques is straightforward: *If the simple linear regression model fits* (possibly after transformation), then it is preferred. The regression approach accomplishes four things: It allows for interpolation, it gives more degrees of freedom for error estimation, it gives smaller standard errors for estimates of the mean responses, and it provides a simpler model.

As an illustration of the increased precision from using regression, Display 8.11 shows estimated means in the insulating fluid example from the separate-means model (i.e., the group averages), and the simple linear regression model (i.e., the fitted values). It also shows three standard errors for the estimated means: the internal standard error of the mean, which is the one-sample standard error (equal to the sample SD over the square root of the sample size); the pooled standard error of the mean, which is the one-way analysis of variance standard error (equal to the pooled SD over the square root of the sample size); and the standard error of the estimated mean response from the regression model (Section 7.4.2). The two nonregression standard errors are comparable in size (but, of course, confidence intervals will be narrower in the second case because of the larger degrees of freedom attached to the pooled SD). The standard errors from the regression model are uniformly smaller, however. Since the mean (at $X = 26$, say) is postulated to be $\beta_0 + \beta_1 26$, all 76 observations are used to estimate this, not just those from group 1.

DISPLAY 8.12

Possible patterns in plots of residuals versus time order of data collection



As a general rule, it is appropriate to find the simplest model—the one with the fewest parameters—that adequately fits the data. This ensures both the most precise answers to the questions of interest and the most straightforward interpretation.

8.6.3 Other Residual Plots for Special Situations

Residuals Versus Time Order

If the data are collected over time (or space), serial correlation may occur. A plot of the residuals versus the time order of data collection may be examined for patterns. Four possibilities are shown in Display 8.12. In part (a) no pattern emerges, and no serial correlation is indicated. In part (b) a linear trend over time can be observed. It may be possible to include time as an additional explanatory variable (using multiple linear regression, as in Chapter 9). The pattern in part (c) shows a positive serial correlation, in which residuals tend to be followed in time by residuals of the same sign and of about the same size. The pattern in part (d) shows a negative serial correlation, where residuals of one sign tend to be followed

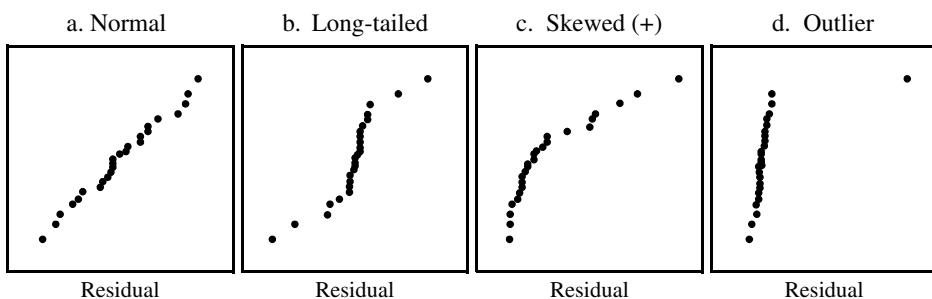
by residuals of the opposite sign. The situations in parts (c) and (d) require time series techniques (Chapter 15).

Normal Probability Plots

The *normal probability plot* is a scatterplot involving the ordered residuals and a set of expected values of an ordered sample of the same size from a standard normal distribution. These expected values are available from statistical theory and are not described further here. Some statistical packages plot the expected values along the y -axis against the ordered residuals along the x -axis (as *The Statistical Sleuth* does), while other packages reverse the axes. Assessing normality visually from a normal plot is easier than from a histogram.

Several normal probability plots for hypothetical data are shown in Display 8.13. When subpopulations have normal distributions, the normal plot is approximately a straight line (a) In a long-tailed distribution (b) residuals in the tails have wider gaps than expected in a normal distribution, with the gaps increasing as one gets further into the tails. A skewed distribution (c) will show wider gaps in one tail and shorter gaps in the other, with a reasonably smooth progression of increase in the longer tail. This is in contrast to a situation with outliers (d) where the bulk of the distribution appears normal with one or a few exceptional cases.

DISPLAY 8.13 Normal probability plots illustrating four distributional patterns



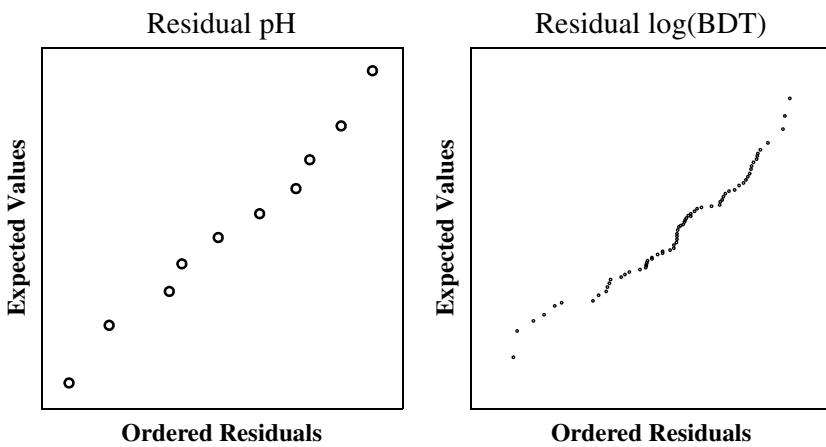
Normal probability plots for the residuals from the meat processing data and the insulating fluid data are shown in Display 8.14. The patterns are close enough to a straight line that the prediction intervals based on the normal assumption should be adequate. (The normal distribution of observations about the regression line is important for the validity of prediction intervals, but not for the validity of estimates, tests, and confidence intervals.)

8.6.4 Planning an Experiment: Balance

Balance means having the same number of experimental units in each treatment group. The insulating fluid data are *unbalanced*, because there are three

DISPLAY 8.14

Normal probability plots of residuals from simple linear regression fits to the meat processing data and the insulating fluid data



observations in the first treatment group, five in the next, and so on. For the several-treatment experiment, balance is generally desirable in providing equal accuracy for all treatment comparisons, but it is not essential. The voltage experiment was designed as unbalanced presumably because of the much greater waiting time for breakdowns at low voltages and the primary interest in voltages between 30 and 36 kV. Balance will play a more important role when the data are cross-classified according to two factors since some simplifying formulas are only appropriate for balanced data and, more importantly, it allows unambiguous decomposition in the analysis of variance.

8.7 SUMMARY

Exploring statistical relationships begins with viewing scatterplots. Nonlinear regressions, nonconstant spread, and outliers can often be identified at this stage. In cases where problems are less apparent, a simple linear regression can be fit tentatively, and the decision about its appropriateness can be based on the residual plot.

When replicate response variables occur at some of the explanatory variable values, it is possible to conduct a formal lack-of-fit F -test. The test is a special case of the extra-sum-of-squares F -test for comparing two models. The models involved are the simple linear regression (reduced) model and the separate-means (full) model.

Insulating Fluid and Species-Area Studies

Scatterplots and residual plots for the insulating fluid data and for the species-area data reveal nonconstant spread and nonlinear regressions, suggesting transformation of the response variable. In both cases, the spread increases as the mean level increases, indicating a logarithmic, square root, or reciprocal transformation. The scatterplot does not indicate which transformation is best. Several may be tried, with the final choice depending on what is appropriate to the scientific context of the study and to the statistical model assumptions.

After a logarithmic transformation of the times to breakdown, a simple linear regression model fits the insulating fluid data well. No evidence (from a residual plot and a lack-of-fit test) indicates lack of fit or (from the normal plot nonnormality) anything but a normal distribution of the residuals. Mean estimation and prediction can proceed from that model, with results back-transformed to the original scale. Other approaches are possible—another sensible analysis of these data assumes the Weibull distribution on the original scale and gives similar results.

To estimate the parameters in the species-area study, both response and explanatory variables are log-transformed. Here the logarithmic transformations to a simple linear regression model are indicated by theoretical model considerations. Weak evidence remains of increasing variability in the residual plot and of long-tailedness in the normal plot. These data should not, however, be used for predictions. With this small sample size, no further action is required, but confidence limits should be described as approximate.

8.8 EXERCISES

Conceptual Exercises

- 1. Island Area and Species Count.** The estimated regression line for the data of Section 8.1.1 is $\hat{\mu}\{\log \text{species} \mid \log \text{area}\} = 1.94 + 0.250 \log(\text{area})$. Show how this estimates that islands of area $0.5A$ have a median number of species that is 16% lower than the median number of species for islands of area A .
- 2. Insulating Fluid.** For the insulating fluid data of Section 8.1.2 explain why the regression analysis allows for statements about the distribution of breakdown times at 27 kV while the one-way analysis of variance does not.
- 3. Big Bang.** In the data set of Section 7.1.1 multiple distances were associated with a few recession velocities. Would it be possible to perform the lack-of-fit test for these data?
- 4. Insulating Fluid.** If the sample correlation coefficient between the square root of breakdown time and voltage is -0.648 , what is R^2 for the regression of square root of breakdown time on voltage?
- 5.** Why can an R^2 close to 1 not be used as evidence that the simple linear regression model is appropriate?
- 6.** A study is made of the stress response exhibited by a sample of 45 adults to rock music played at nine different volume levels (five adults at each level). What is the difference between using the

volume as an explanatory variable in a simple linear regression model and using the volume level as a group designator in a one-way classification model?

7. In a study where four levels of a magnetic resonance imaging (MRI) agent are each given to 3 cancer patients (so there are 12 patients in all), the response is a measure of the degree of seizure activity (an unpleasant side effect). The F -test for lack of fit to the simple linear regression model with $X = \text{agent level}$ has a p -value of 0.0082. The t -tools estimate that the effect of increasing the level of the MRI agent by 1 mg/cm^2 is to increase the level of seizure activity by 2.6 units (95% confidence interval from 1.8 to 3.4 units). (a) How should the latter inference be interpreted? (b) How many degrees of freedom are there for (i) the within-group variation? (ii) the lack-of-fit variation?

8. Insulating Fluid. Why would it be of interest to know whether batches of insulating fluid were *randomly* assigned to the different voltage levels?

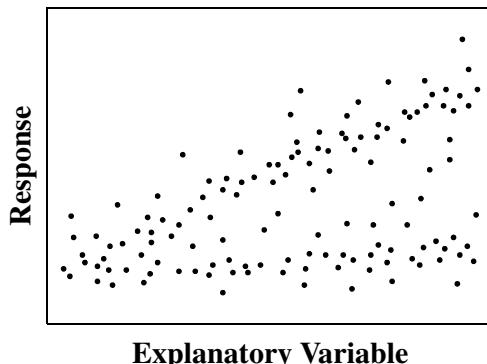
9. Suppose the (Y, X) pairs are: $(5,1)$, $(3,2)$, $(4,3)$, $(2,4)$, $(3,5)$, and $(1,6)$. Would the least squares fit to these data be much different from the least squares fit to the same data with the first pair replaced by $(15,1)$?

10. (a) What assumptions are used for exact justification of tests and confidence intervals for the slope and intercept in simple regression? (b) Are any of these assumptions relatively unimportant?

11. Suppose you had data on pairs (Y, X) which gave the scatterplot shown in Display 8.15. How would you approach the analysis?

DISPLAY 8.15

Scatterplot for Exercise 11



12. What is the technical difficulty with using the separate-means model as a basis for the lack-of-fit F -test when there are no replicate responses?

13. Researchers at a university wish to estimate the effect of class size on course comprehension. An intermediate course in statistics can be taught to classes of any size between 25 and 185 students, and four instructors are available. Suppose the researchers truly believe that the average course comprehension, measured by the average of student scores on a standardized test, is indeed a straight line in class sizes over the range from 25 to 185. What four class sizes should be used in the experiment? Why?

14. Insulating Fluid. Which would you use to predict the log breakdown time for a batch of insulating fluid which is to be put on test at 30 kV: the regression estimate of the mean at 30 kV or the average from the batches that were tested at 30 kV? Why?

Computational Exercises

15. Island Size and Species. (a) Draw a scatterplot of the (untransformed) number of species on the (untransformed) area of the island (Display 8.2, top). (b) Fit the simple linear regression of number of species on area and obtain a residual plot. (c) What features in the two plots indicate a need for transformation?

16. Meat Processing. The data in Display 7.3 are a subset of the complete data on postmortem pH in 12 steer carcasses. (Data from J. R. Schwenke and G. A. Milliken, “On the Calibration Problem Extended to Nonlinear Models,” *Biometrics* 47(2) (1991): 563–74). Once again, the purpose is to determine how much time after slaughter is needed to ensure that the pH reaches 6.0. In Chapter 7 the simple linear regression of pH on log(Hour) was fit to the first 10 carcasses only. Refit the model with all 12 carcasses (data given in Display 8.16). (a) Assess lack of fit using a residual plot. (b) Assess lack of fit using the lack-of-fit *F*-test. (c) The inappropriateness of the simple linear regression model can be remedied by dropping the last two carcasses. Is there any justification for doing so? (*Hint:* In order to answer the question of interest, what range of X ’s appears to be important?)

DISPLAY 8.16

pH of steer carcasses 1 to 24 hours after slaughter

Animal Number:	1	2	3	4	5	6	7	8	9	10	11	12
Processing Hour:	1	1	2	2	4	4	6	6	8	8	24	24
pH:	7.02	6.93	6.42	6.51	6.07	5.99	5.59	5.80	5.51	5.36	5.30	5.47

17. Biological Pest Control. In a study of the effectiveness of biological control of the exotic weed tansy ragwort, researchers manipulated the exposure to the ragwort flea beetle on 15 plots that had been planted with a high density of ragwort. Harvesting the plots the next season, they measured the average dry mass of ragwort remaining (grams/plant) and the flea beetle load (beetles/gram of ragwort dry mass) to see if the ragwort plants in plots with high flea beetle loads were smaller as a result of herbivory by the beetles. (Data from P. McEvoy and C. Cox, “Successful Biological Control of Ragwort, *Senecio jacobaea*, by Introduced Insects in Oregon,” *Ecological Applications* 1(4) (1991): 430–42. The data in Display 8.17 were read from McEvoy and Cox, Figure #2.)

DISPLAY 8.17

Dry mass of ragwort weed on 15 plots exposed to flea beetles

Plot #:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Flea beetle load:	12.2	14.6	15.8	25.3	38.6	76.4	163	182	415	446	628	377	770	1,446	1,012
Ragwort mass:	18.2	17.5	7.22	30.6	6.66	6.14	5.21	0.502	0.611	0.630	0.427	0.011	0.012	0.006	0.002

- (a) Use scatterplots of the raw data, along with trial and error, to determine transformations of Y = Ragwort dry mass and of X = Flea beetle load that will produce an approximate linear relationship.
- (b) Fit a linear regression model on the transformed scale; calculate residuals and fitted values.
- (c) Look at the residual plot. Do you want to try other transformations? What do you suggest?

18. Distance and Order from Sun. Reconsider the planetary distance and order from sun data in Exercise 7.21. Fit a regression model to the data that includes the asteroid belt and fill in the blanks in this conclusion: Aside from some random variation, the distance to the sun increases by ____% with each consecutive planet number (95% confidence interval: ____ to ____% increase).

19. Pollen Removal. Reconsider the pollen removal data of Exercise 3.27 and the regression of pollen removed on time spent on flower, for the bumblebee queens only. (a) What problems are evident in the residual plot? (b) Do log transformations of Y or X help any? (c) Try fitting the regression only for those times less than 31 seconds (i.e., excluding the two longest times). Does this fit better? (*Note:* If the linear regression fits for a restricted range of the X 's, it is acceptable to fit the model with all the other X 's excluded and to report the range of X 's for which the model holds.)

20. Quantifying Evidence for Outlierness. In a special election to fill a Pennsylvania State Senate seat in 1993, the Democratic candidate, William Stinson, received 19,127 machine-counted votes and the Republican, Bruce Marks, received 19,691 (i.e., 564 more votes than the Democrat). In addition, however, Stinson received 1,396 absentee ballots and Marks received only 371, so the total tally showed the Democrat, Stinson, winning by 461 votes. The large disparity between the machine-counted and absentee ratios, and the resulting reversal of the outcome due to the absentee ballots, sparked concern about possible illegal influence on the absentee votes. Investigators reviewed data on disparities between machine and absentee votes in prior Pennsylvania State Senate elections to see whether the 1993 disparity was larger than could be explained by historical variation. Display 8.18 shows the data in the form of percentage of absentee and machine-counted ballots cast for the Democratic candidate. The task is to clarify the unusualness of the Democratic absentee percentage in the disputed election. (a) Draw a scatterplot of Democratic percentage of absentee ballots versus Democratic percentage of machine-counted ballots. Use a separate plotting symbol to highlight the disputed election. (b) Fit the simple linear regression of absentee percentage on machine-count percentage, *excluding* the disputed election. Draw this line on the scatterplot. Also include a 95% prediction band. What does this plot reveal about the unusualness of the absentee percentage in the disputed election? (c) Find the prediction and standard error of prediction from this fit if the machine-count percentage is 49.3 (as it is for the disputed election). How many estimated standard deviations is the observed absentee percentage, 79.0, from this predicted value? Compare this answer to a t -distribution (with degrees of freedom equal to the residual degrees of freedom in the regression fit) to obtain a p -value. (d) **Outliers and data snooping.** The p -value in (c) makes sense if the investigation into the 1993 election was prompted by some other reason. Since it was prompted because the absentee percentage seemed too high, however, the p -value in (c) should be adjusted for data snooping. Adjust the p -value with a Bonferroni correction to account for all 22

DISPLAY 8.18

Partial listing of 22 Pennsylvania state senate election results, collected to explore the peculiarity noticed in the 1993 election between William Stinson and Bruce Marks (election number 22 in the data set). The full data set includes the year of the election, the senate district, the number of absentee ballots cast for the Democratic candidate and for the Republican candidate, and the number of machine-counted ballots cast for the Democratic candidate and for the Republican candidate. Also shown are the percentages of absentee and machine ballots cast for the Democratic candidate.

Election	Year	District	DemPctOfAbsenteeVotes	DemPctOfMachineVotes	Disputed
1	82	D2	72.9	69.1	no
2	82	D4	65.6	60.9	no
3	82	D8	74.6	80.8	no
4	84	D1	64.0	60.0	no
5	84	D3	83.3	92.4	no
...					
22	93	D2	79.0	49.3	yes

residuals that could have been similarly considered. (Data from Orley Ashenfelter, 1994. Report on Expected Absentee Ballots. Typescript. Department of Economics, Princeton University. See also Simon Jackman (2011). pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory, Stanford University. Department of Political Science, Stanford University. Stanford, California. R package version 1.03.10. URL <http://pscl.stanford.edu/>)

21. Fish Preferences. Reconsider Case Study 2 in Chapter 6, the study of female preferences among platyfish. Fit the full model in which the mean preference for the yellowtailed male is possibly different for each male pair. Construct a normal probability plot of the residuals and a residual plot. If these suggest a transformation, make it and repeat the analysis including the linear contrast measuring the association of preference with male body size. If they suggest an outlier problem, use the inclusion/exclusion procedure to determine whether the outlying case(s) change(s) the answer to the questions of interest. Also, identify the outlying case(s) and suggest why it might be a true outlier.

Data Problems

22. Ecosystem Decay. As an introduction to their study on the effect of Amazon forest clearing (data from T. E. Lovejoy, J. M. Rankin, R. O. Bierregaard, Jr., K. S. Brown, Jr., L. H. Emmons, and M. E. Van der Woot, “Ecosystem Decay of Amazon Forest Remnants,” in M. H. Nitecki, ed., *Extinctions*, Chicago: University of Chicago Press, 1984) the researchers stated: “fragmentation of once continuous wild areas is a major way in which people are altering the landscape and biology of the planet.” Their study takes advantage of a Brazilian requirement that 50% of the land in any development project remain in forest and tree cover. As a consequence of this requirement, “islands” of forest of various sizes remain in otherwise cleared areas. The data in Display 8.19 are the number of butterfly species in 16 such islands. Summarize the role of area in the distribution of number of butterfly species. Write a brief statistical report including a summary of statistical findings, a graphical display, and a section detailing the methods used to answer the questions of interest.

DISPLAY 8.19

Forest patch area (hectares) and number of butterfly species found

Reserve	Area	Species	Reserve	Area	Species
1	1	14	9	10	33
2	1	50	10	10	53
3	1	55	11	10	50
4	1	34	12	100	110
5	1	40	13	100	70
6	1	57	14	100	119
7	10	43	15	100	60
8	10	103	16	1,000	145

23. Wine Consumption and Heart Disease. The data in Display 8.20 are the average wine consumption rates (in liters per person) and number of ischemic heart disease deaths (per 1,000 men aged 55 to 64 years old) for 18 industrialized countries. (Data from A. S. St. Leger, A. L. Cochrane, and F. Moore, “Factors Associated with Cardiac Mortality in Developed Countries with Particular Reference to the Consumption of Wine,” *Lancet* (June 16, 1979): 1017–20.) Do these data suggest that the heart disease death rate is associated with average wine consumption? If so, how can that

DISPLAY 8.20

First five rows of a data set with wine consumption (liters per person per year) and heart disease mortality rates (deaths per 1,000) in 18 countries

Country	Wine consumption	Heart disease mortality
Norway	2.8	6.2
Scotland	3.2	9.0
England	3.2	7.1
Ireland	3.4	6.8
Finland	4.3	10.2

relationship be described? Do any countries have substantially higher or lower death rates than others with similar wine consumption rates? Analyze the data and write a brief statistical report that includes a summary of statistical findings, a graphical display, and a section detailing the methods used to answer the questions of interest.

24. Respiratory Rates for Children. A high respiratory rate is a potential diagnostic indicator of respiratory infection in children. To judge whether a respiratory rate is truly “high,” however, a physician must have a clear picture of the distribution of *normal* respiratory rates. To this end, Italian researchers measured the respiratory rates of 618 children between the ages of 15 days and 3 years. Display 8.21 shows a few rows of the data set. Analyze the data and provide a statistical summary. Include a useful plot or chart that a physician could use to assess a normal range of respiratory rate for children of any age between 0 and 3. (Data read from a graph in Rusconi et al., “Reference Values for Respiratory Rate in the First 3 Years of Life,” *Pediatrics*, 94 (1994): 350–55.)

DISPLAY 8.21

Partial listing of data on ages (months) and respiratory rates (breaths per minute) for 618 children

Child	1	2	3	4	5	6	...	618
Age:	0.1	0.2	0.3	0.3	0.3	0.4	...	36.0
Rate:	53	38	58	52	42	62	...	31

25. The Dramatic U.S. Presidential Election of 2000. The U.S. presidential election of November 7, 2000, was one of the closest in history. As returns were counted on election night it became clear that the outcome in the state of Florida would determine the next president. At one point in the evening, television networks projected that the state was carried by the Democratic nominee, Al Gore, but a retraction of the projection followed a few hours later. Then, early in the morning of November 8, the networks projected that the Republican nominee, George W. Bush, had carried Florida and won the presidency. Gore called Bush to concede. While en route to his concession speech, though, the Florida count changed rapidly in his favor. The networks once again reversed their projection, and Gore called Bush to retract his concession. When the roughly 6 million Florida votes had been counted, Bush was shown to be leading by only 1,738, and the narrow margin triggered an automatic recount. The recount, completed in the evening of November 9, showed Bush’s lead to be less than 400.

Meanwhile, angry Democratic voters in Palm Beach County complained that a confusing “butterfly” lay-out ballot caused them to accidentally vote for the Reform Party candidate Pat Buchanan instead of Gore. The ballot, as illustrated in Display 8.22, listed presidential candidates on both a left-hand and a right-hand page. Voters were to register their vote by punching the circle

DISPLAY 8.22 Confusing ballot in Palm Beach County, Florida

ELECTORS for PRESIDENT and VICE PRESIDENT	(REPUBLICAN) GEORGE W. BUSH-President DICK CHENEY-Vice President	3 ►	<input type="radio"/>
	(DEMOCRATIC) AL GORE-President JOE LIBERMAN-Vice President	5 ►	<input type="radio"/>
	(LIBERTARIAN) HARRY BROWNE-President ART OLIVER-Vice President	7 ►	<input type="radio"/>
		9 ►	<input type="radio"/>
		11 ►	<input type="radio"/>
			<input type="radio"/>

corresponding to their choice, from the column of circles between the pages. It was suspected that since Bush's name was listed first on the left-hand page, Bush voters likely selected the first circle. Since Gore's name was listed second on the left-hand side, many voters—who already knew who they wished to vote for—did not bother examining the right-hand side and consequently selected the second circle in the column; the one actually corresponding to Buchanan. Two pieces of evidence supported this claim: Buchanan had an unusually high percentage of the vote in that county, and an unusually large number of ballots (19,000) were discarded because voters had marked two circles (possibly by inadvertently voting for Buchanan and then trying to correct the mistake by then voting for Gore).

Display 8.23 shows the first few rows of a data set containing the numbers of votes for Buchanan and Bush in all 67 counties in Florida. What evidence is there in the scatterplot of Display 8.24 that Buchanan received more votes than expected in Palm Beach County? Analyze the data without Palm Beach County results to obtain an equation for predicting Buchanan votes from Bush votes. Obtain a 95% prediction interval for the number of Buchanan votes in Palm Beach from this result—assuming the relationship is the same in this county as in the others. If it is assumed that Buchanan's actual count contains a number of votes intended for Gore, what can be said about the likely size of this number from the prediction interval? (Consider transformation.)

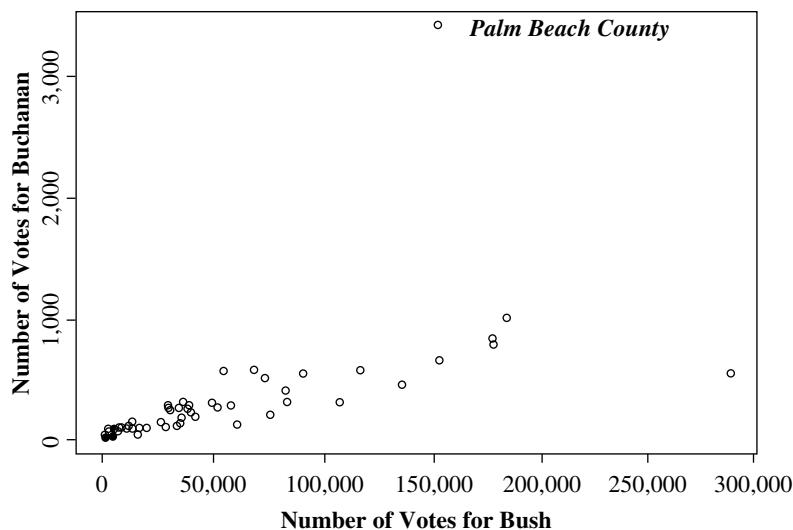
DISPLAY 8.23 Votes for Bush and Buchanan in all Florida counties (first 5 of 67 rows)

County	Bush votes	Buchanan votes
Alachua	34,062	262
Baker	5,610	73
Bay	38,637	248
Bradford	5,413	65
Brevard	115,185	570

- 26. Kleiber's Law.** Display 8.25 shows the first five rows of a data set with average mass, metabolic rate, and average lifespan for 95 species of mammals. (From A. T. Atanasov, "The Linear Allo-metric Relationship Between Total Metabolic Energy per Life Span and Body Mass of Mammals," *Biosystems* 90 (2007): 224–33.) Kleiber's law states that the metabolic rate of an animal species, on

DISPLAY 8.24

Votes for George W. Bush and Pat Buchanan in all Florida counties

**DISPLAY 8.25**

Average mass (kg), average basal metabolic rate (kJ per day), and lifespan (years) for 95 mammal species; first 5 of 95 rows

CommonName	Species	Mass	Metab	Life
Echidna	<i>Tachiglossus aculeatus</i>	2.50E+00	3.02E+02	14
Long-beaked echidna	<i>Zaglossus bruijni</i>	1.03E+01	5.94E+02	20
Platypus	<i>Ornithorhynchus anatinus</i>	1.30E+00	2.29E+02	9
Opossum	<i>Lutreolina crassicaudata</i>	8.12E-01	1.96E+02	5
South American opossum	<i>Didelphis marsupialis</i>	1.33E+00	2.99E+02	6

average, is proportional to its mass raised to the power of 3/4. Judge the adequacy of this theory with these data.

27. Metabolic Rate and Lifespan. It has been suggested that metabolic rate is one of the best single predictors of species lifespan. Analyze the data in Exercise 26 to describe an equation for predicting mammal lifespan from metabolic rate. Also provide a measure of the amount of variation in the distribution of mammal lifespans that can be explained by metabolic rate.

28. IQ, Education, and Future Income. Display 8.26 is a partial listing of a data set with IQ scores in 1981, years of education completed by 2006, and annual income in 2005 for 2,584 Americans who were selected for the National Longitudinal Study of Youth in 1979 (NLSY79), who were re-interviewed in 2006, and who had paying jobs in 2005. (See Exercises 2.22 and 3.30 for a more detailed description of the survey.) (a) Describe the distribution of 2005 income as a function of IQ test score. What percentage of variation in the distribution is explained by the regression? (b) Describe the distribution of 2005 income as a function of years of education. What percentage of variation in the distribution is explained by the regression?

DISPLAY 8.26

IQ test score from 1981 (AFQT—armed forces qualifying test score), number of years of education, and annual income in 2005 (dollars) for 2,584 Americans in the NLSY79 survey; first 5 of 2,584 rows

Subject	AFQT	Educ	Income2005
2	6.841	12	5,500
6	99.393	16	65,000
7	47.412	12	19,000
8	44.022	14	36,000
9	59.683	14	65,000

DISPLAY 8.27

Autism prevalence per 10,000 ten-year olds in each of five years

Year	Prevalence
1992	3.5
1994	5.3
1996	7.8
1998	11.8
2000	18.3

29. Autism Rates. Display 8.27 shows the prevalence of autism per 10,000 ten-year old children in the United States in each of five years. Analyze the data to describe the change in the distribution of autism prevalence per year in this time period. (Data from C. J. Newschaffer, M. D. Falb, and J. G. Gurney, “National Autism Prevalence Trends From United States Special Education Data,” *Pediatrics*, 115 (2005): e277–e282.)

Answers to Conceptual Exercises

- Median {species|area} = $\exp(1.94)\text{area}^{0.250}$. So Median{species | 0.5 area}/Median{species | area} = $0.5^{0.250} = 0.84$. Thus Median{species | 0.5 area} = 0.84 Median{species | area} and, finally, [Median{species | area} – Median{species | 0.5 area}]/Median{species | area} = 1 – 0.84 = 0.16.
- The ANOVA model states that there are seven means, one for each voltage level tested, but does not describe any relation between mean and voltage. The regression model establishes a pattern between mean log breakdown time and voltage, for all voltages in the range of 26 to 38 kV.
- Yes. (*Note:* The multiple observations occurred because of rounding, so they do not represent repeated draws from the same distribution. Nevertheless, they are near replicates, at least, and can be used as such for the lack-of-fit test.)
- R^2 = square of correlation coefficient = $(-0.648)^2 = 0.420$.
- Although a high R^2 reflects a strong degree of linear association, this linear association may well be accompanied by curvature (and by nonconstant variance).
- In the simple linear regression model, the nine mean stress levels lie on a straight line against volume. In the one-way classification (the separate-means model) the mean stress levels may or may not lie on the straight line—their values are not restricted.

7. (a) If the data do not fit the model, then the parameters of the model are not adequate descriptive summaries. No inference should be drawn, at least until a better model is found. (b) (i) 8; (ii) 2.
8. Even in laboratory circumstances, confounding variables are possible. If, for example, batches are spooned from a large container that has density stratification, then assigning consecutive batches to the lowest voltage, then the next lowest, and so on, will confound fluid density with voltage. Randomization never hurts, and usually helps.
9. Yes, very much so. The least squares method is not resistant to the effects of outliers.
10. (a) linearity, constant variance, normality, independence. (b) normality.
11. It may appear that this is a case where the spread of Y increases as the mean of Y does, so a simple transformation may be in order. This will not produce good results. Notice that the average Y at a large X lies in a no-man's land with few observations. Looking at the distributions in strips, you will see two separate groups in each distribution. Look for some important characteristic that separates the data into two groups, then build a separate regression model for each group.
12. In fitting the separate-means model, all residuals are zero. The denominator of the F -statistic is zero, so the F -statistic is not defined.
13. Put two classes at 25 students and two at 185 students. This makes the standard deviation in the sampling distribution of the slope parameter as small as possible, given the constraints of the situation. (But it gives no information on lack of fit; check the degrees of freedom.)
14. The regression estimate. It is more precise (see Display 8.11).

Multiple Regression

Multiple regression analysis is one of the most widely used statistical tools, and for good reason: It is remarkably effective for answering questions involving many variables. Although more difficult to visualize than simple regression, multiple regression is a straightforward extension. It models the mean of a response variable as a function of *several* explanatory variables.

Many issues, tools, and strategies are associated with multiple regression analysis, as discussed in this and the next three chapters. This chapter focuses on the meaning of the regression model and strategies for analysis. The details of estimation and inferential tools are deliberately postponed until the next chapter so that the student may concentrate first on understanding regression coefficients and the types of data structures that may be analyzed with multiple regression analysis.

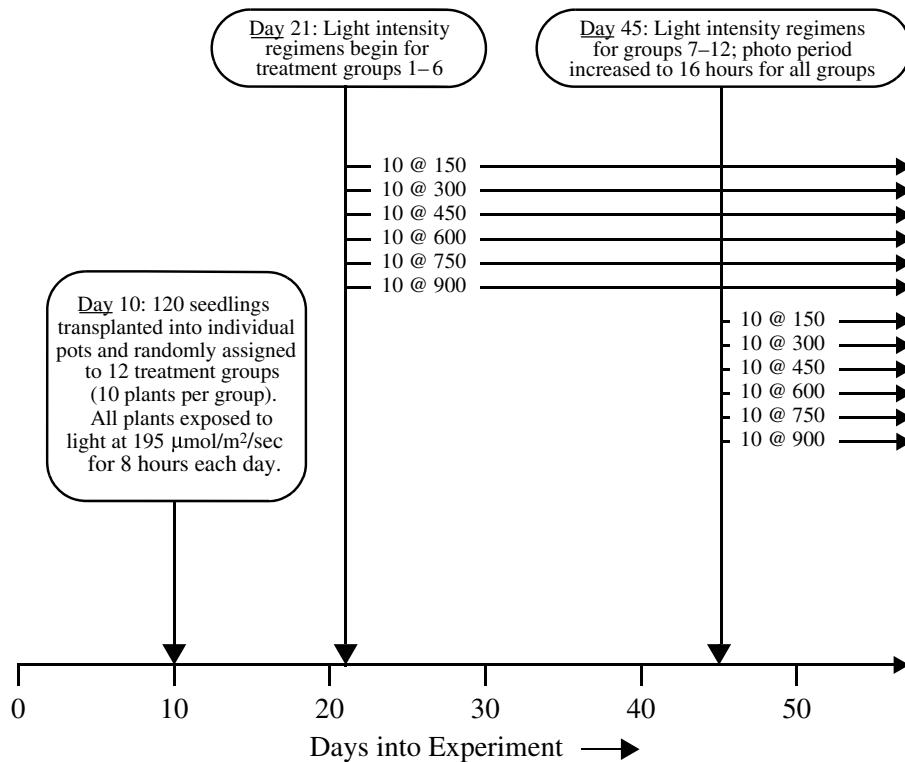
9.1 CASE STUDIES

9.1.1 Effects of Light on Meadowfoam Flowering—A Randomized Experiment

Meadowfoam (*Limnanthes alba*) is a small plant found growing in moist meadows of the U.S. Pacific Northwest. It has been domesticated at Oregon State University for its seed oil, which is unique among vegetable oils for its long carbon strings. Like the oil from sperm whales, it is nongreasy and highly stable.

Researchers reported the results from one study in a series designed to find out how to elevate meadowfoam production to a profitable crop. In a controlled growth chamber, they focused on the effects of two light-related factors: light intensity, at the six levels of 150, 300, 450, 600, 750, and 900 $\mu\text{mol}/\text{m}^2/\text{sec}$; and the timing of the onset of the light treatment, either at photoperiodic floral induction (PFI)—the time at which the photo period was increased from 8 to 16 hours per day to induce flowering—or 24 days before PFI. The experimental design is depicted in Display 9.1. (Data from M. Seddigh and G. D. Jolliff, “Light Intensity Effects on Meadowfoam Growth and Flowering,” *Crop Science* 34 (1994): 497–503.)

DISPLAY 9.1 Time line for light variation experiment on meadowfoam



The design consists of 12 treatment groups—the six light intensities at each of the two timing levels. Ten seedlings were randomly assigned to each treatment group. The number of flowers per plant is the primary measure of production, and it was measured by averaging the numbers of flowers produced by the 10 seedlings in each group.

The entire experiment was repeated. No difference was found between the first and second runs of the experiment so, as a suitable starting point for the analysis in this chapter, Display 9.2 presents the two results as replicates. The two observations in each cell of the table are thought of as independent replicates under the specified conditions. What are the effects of differing light intensity levels? What is the effect of the timing? Does the effect of intensity depend on the timing?

		Light Intensity ($\mu\text{mol}/\text{m}^2/\text{sec}$)					
		150	300	450	600	750	900
Timing	Late (at PFI)	62.3 77.4	55.3 54.2	49.6 61.9	39.4 45.7	31.3 44.9	36.8 41.9
	Early (24 days before PFI)	77.8 75.6	69.1 78.0	57.0 71.1	62.9 52.2	60.3 45.6	52.6 44.4

Statistical Conclusion

Display 9.3 shows the fit of a multiple linear regression model that specifies parallel regression lines for the mean number of flowers as functions of light intensity. Increasing light intensity decreased the mean number of flowers per plant by an estimated 4.0 flowers per plant per $100 \mu\text{mol}/\text{m}^2/\text{sec}$ (95% confidence interval from 3.0 to 5.1). Beginning the light treatments 24 days prior to PFI increased the mean numbers of flowers by an estimated 12.2 flowers per plant (95% confidence interval from 6.7 to 17.6). The data provide no evidence that the effect of light intensity depends on the timing of its initiation (two-sided p -value = 0.91, from a t -test for interaction, 20 degrees of freedom).

Scope of Inference

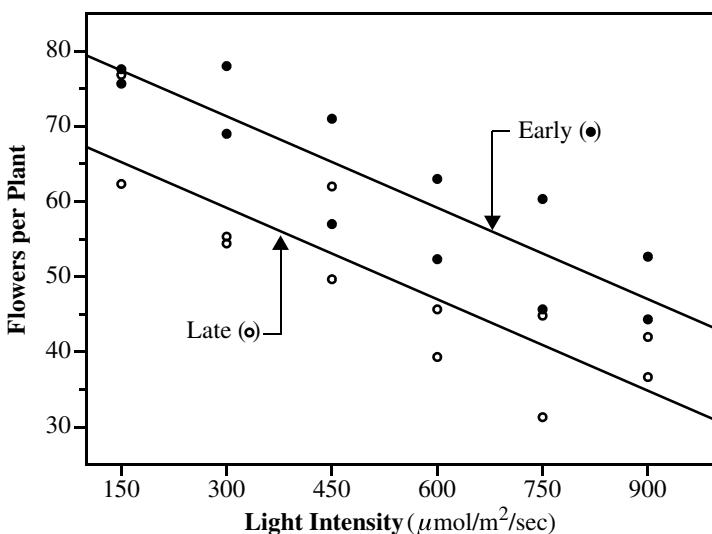
The researchers can infer that the effects above were caused by the light intensity and timing manipulations, because this was a randomized experiment.

9.1.2 Why Do Some Mammals Have Large Brains for Their Size?—An Observational Study

Evolutionary biologists are keenly interested in the characteristics that enable a species to withstand the selective mechanisms of evolution. An interesting variable

DISPLAY 9.3

The average number of flowers per plant versus the applied light intensity for each of the 12 experimental units in the meadowfoam study, with different plotting symbols for units with early (24 days prior to PFI) and late (at PFI) commencement of lighting regimen



in this respect is brain size. One might expect that bigger brains are better, but certain penalties seem to be associated with large brains, such as the need for longer pregnancies and fewer offspring. Although the individual members of the large-brained species may have more chance of surviving, the benefits for the species must be good enough to compensate for these penalties. To shed some light on this issue, it is helpful to determine exactly which characteristics are associated with large brains, after getting the effect of body size out of the way.

The data in Display 9.4 are the average values of brain weight, body weight, gestation lengths (length of pregnancy), and litter size for 96 species of mammals. (Data from G. A. Sacher and E. F. Staffeldt, "Relation of Gestation Time to Brain Weight for Placental Mammals; Implications for the Theory of Vertebrate Growth," *American Naturalist*, 108 (1974): 593–613. The common names for the species correspond to the Latin names given in the original paper, and those followed by a Roman numeral indicate subspecies.) Since brain size is obviously related to body size, the question of interest is this: Which, if any, variables are associated with brain size, after accounting for body size?

Statistical Conclusion

The data provide convincing evidence that brain weight was associated with either gestation length or litter size, even after accounting for the effect of body weight (p -value < 0.0001 ; extra sum of squares F -test). There was strong evidence that litter size was associated with brain weight after accounting for body weight and gestation (two-sided p -value = 0.0089) and that gestation period was associated

DISPLAY 9.4

Average values of brain weight, body weight, gestation length, and litter size in 96 species of mammal

Species	Litter size	Gestation period (days)	Body weight (kilograms)	Brain weight (grams)	Species	Litter size	Gestation period (days)	Body weight (kilograms)	Brain weight (grams)
Quokka	17.5	3.5	26	1.0	A couchis	9.9	0.78	98	1.2
Hedgehog	3.50	0.93	34	4.6	Chinchilla	5.25	0.43	110	2.0
Tree shrew	3.15	0.15	46	3.0	Nutria	23	5.0	132	5.5
Elephant shrew I	1.14	0.049	51	1.5	Dolphin	1,600	160	360	1.0
Elephant shrew II	1.37	0.064	46	1.5	Porpoise	537	56	270	1.0
Lemur	22	2.1	135	1.0	Dog	70.2	8.5	63	4.0
Slow loris	12.8	1.2	90	1.2	Red fox	48	6.0	52	4.0
Bush baby	9.9	0.7	135	1.0	Gray fox	37.3	3.8	63	3.7
Howler monkey	54	7.7	139	1.0	Bat-eared fox	28.5	3.2	65	4.0
Ring-tail monkey	73	3.7	180	1.0	Grizzly bear	400	250	219	2.3
Spider monkey I	114	9.1	140	1.0	Beaked whale	500	250	240	1.8
Spider monkey II	109	7.7	140	1.0	Raccoon	41.6	5.3	63	3.5
Gentle lemur	7.8	0.22	145	2.0	Kinkajou	31.2	2.0	77	1.1
Rhesus monkey I	84.6	6.0	175	1.0	Badger	53	6.0	60	2.2
Rhesus monkey II	107	8.7	165	1.1	Domestic cat	28.4	2.5	63	4.0
Hamadryas baboon	183	21	180	1.0	Lynx	75	12	60	2.5
Western baboon	179	32	180	1.0	Leopard	157	46	92	2.5
Vervet guenon	67	4.6	195	1.0	Lion	260	180	108	3.0
Leaf monkey	65.5	5.8	168	1.0	Tiger	302	210	104	3.0
White-handed gibbon	102	5.5	210	1.0	Fur seal	355	250	254	1.0
Orangutan	343	37	270	1.0	Sea lion	363	100	343	1.0
Chimpanzee	360	45	230	1.0	Harp seal	442	110	240	1.0
Gorilla	406	140	265	1.0	Weddell seal	550	400	310	1.0
Human being	1,300	65	270	1.0	African Elephant	4,480	2,800	655	1.0
Long-nosed armadillo	12	3.7	120	4.0	Hyrax	20.5	3.8	225	2.4
Aardvark	9.6	2.2	31	5.0	Horse	712	480	330	1.0
Jack rabbit	13.3	2.9	41	2.5	Tapir	250	230	390	1.0
Tree squirrel	6.23	0.33	38	3.0	Wild boar	185	150	120	4.0
Flying squirrel	1.89	0.052	40	3.1	Domestic pig	180	190	115	8.0
Canadian beaver	40	20	128	2.9	Hippopotamus	590	1,400	240	1.0
Beaver	45	25	128	4.0	Pygmy hippopotamus	260	150	205	1.0
Deer mouse I	0.68	0.027	23	3.7	Llama	225	93	330	1.0
Deer mouse II	0.63	0.026	23	5.0	Vicuna	198	45	300	1.1
Deer mouse III	0.52	0.017	24	5.0	Barking deer	124	16	183	1.1
Deer mouse IV	0.69	0.024	24	5.0	Fallow deer	223	80	240	1.0
Hamster I	0.67	0.036	21	4.6	Axis deer	219	89	218	1.0
Hamster II	1.12	0.13	16	6.3	Red deer	435	200	255	1.0
Pygmy gerbil	1.04	0.065	21	4.0	Elk	365	120	235	1.0
Rat I	0.72	0.05	23	7.3	Sambar	383	120	246	1.1
Rat II	2.38	0.34	21	8.0	Caribou	288	110	225	1.0
House mouse	0.45	0.024	19	5.0	Eland	480	560	255	1.0
Hopping mouse	1.18	0.15	27	5.6	Yak	334	250	255	1.0
Porcupine I	37	11	112	1.2	Cattle	456	520	280	1.0
Porcupine II	37	14	112	1.2	Duiker	93	13	120	1.0
Porcupine III	24	6.6	113	1.0	Blackbuck Antelope	200	39	180	1.0
Guinea pig	4.28	0.97	67	2.6	Barbary sheep	210	66	158	1.2
Capybara	76	30	123	3.0	Domestic sheep	125	49	150	2.4
Agoutis	20.3	2.8	104	1.3	Domestic goat	106	30	151	2.0

with brain weight after accounting for body weight and litter size (two-sided p -value = 0.0038).

Scope of Inference

As suggestive as the findings may be, inferences that go beyond these data are unwise. The data were summarized from available studies and cannot be representative of any wider population. As usual, no causal interpretation can be made from these observational data.

9.2 REGRESSION COEFFICIENTS

9.2.1 The Multiple Linear Regression Model

Regression

The *regression of Y on X_1 and X_2* is a rule, such as an equation, that describes the mean of the distribution of a response variable (Y) for particular values of explanatory variables (X_1 and X_2 , say). For example, the regression of the response variable $Y = \text{flowers}$ on $X_1 = \text{light}$ intensity and $X_2 = \text{time}$ of light manipulation specifies how the mean number of flowers per plant depends on the levels of light intensity and timing. (The italicized words specify variable names.)

The symbol for the regression is $\mu\{\text{flowers} \mid \text{light}, \text{time}\}$, which is read as the “mean number of flowers, as a function of light intensity and timing.” In terms of a generic response variable, Y , and two explanatory variables, X_1 and X_2 :

The regression of Y on X_1 and X_2 is $\mu\{Y|X_1, X_2\}$.

With specific values for X_1 and X_2 inserted, the same expression may be read somewhat differently. For example, $\mu\{\text{flowers} \mid \text{light} = 300, \text{time} = 24\}$ is read as “the mean number of flowers when light intensity is $300 \mu\text{mol}/\text{m}^2/\text{sec}$ and time is 24 days prior to PFI.”

Multiple Regression Models

In *multiple regression* there is a single response variable and *multiple explanatory variables*. The regression rule giving the mean response for each combination of explanatory variables will not be very helpful if it is a huge list or a complicated function. Furthermore, it is usually unwise to think that there is some exact, discoverable regression equation. Many possible *models* are available, however, for describing the regression. Although they should not be thought of as the truth, one or two models may adequately approximate the mean of the response as a function of the explanatory variables, and conveniently allow for the questions of interest to be investigated.

Multiple Linear Regression Model

One family of models is particularly easy to deal with and works for the majority of regression problems—the family of linear regression models. The term *linear* means linear in regression coefficients, as in the following examples:

Examples of Multiple Linear Regression Models

$$\begin{aligned}\mu\{Y|X_1, X_2\} &= \beta_0 + \beta_1 X_1 + \beta_2 X_2, \\ \mu\{Y|X_1\} &= \beta_0 + \beta_1 X_1 + \beta_2 X_1^2, \\ \mu\{Y|X_1, X_2\} &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2, \\ \mu\{Y|X_1, X_2\} &= \beta_0 + \beta_1 \log(X_1) + \beta_2 \log(X_2).\end{aligned}$$

The extension to more than two explanatory variables is straightforward. Notice that the first term in the examples is β_0 (which multiplies the trivial function, 1). This *constant term* appears in multiple linear regression models unless a specific reason for excluding it exists.

Multiple Regression Model with Constant Variance

The ideal regression model includes an assumption of constant variation. For the meadowfoam example, the assumption states that

$$\text{Var}\{\text{flowers} | \text{light}, \text{time}\} = \sigma^2.$$

The notation $\text{Var}\{\text{flowers} | \text{light}, \text{time}\}$ is read as “the variance of the numbers of flowers, as a function of light and time.” The right-hand side of the equation asserts that this variance is constant—the same for all values of *light* and *time*. As in previous chapters, the constant variance assumption is important for two reasons: (1) the regression interpretation is more straightforward when the explanatory variables are only associated with the mean of the response distribution and not other characteristics of it (as described in Section 4.3.2 for the two-sample problem); and (2) the assumption justifies the standard inferential tools, which are presented in the next chapter.

9.2.2 Interpretation of Regression Coefficients

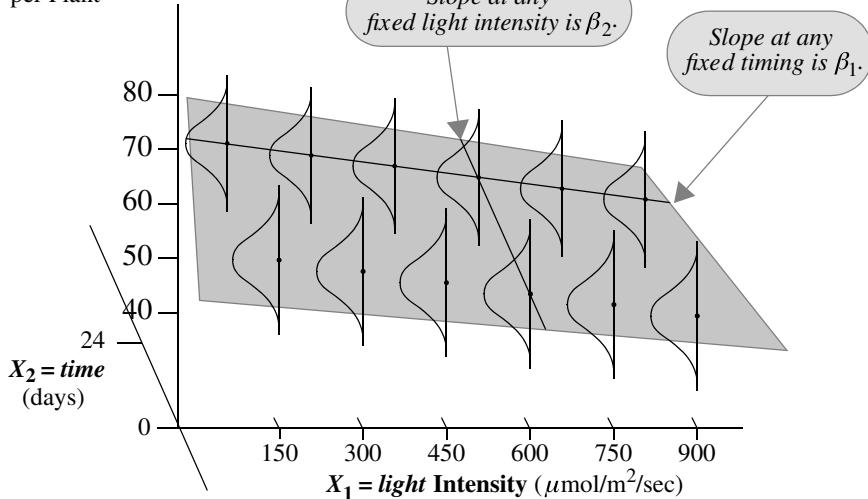
Regression analysis involves finding a good fitting model for the response mean, wording the questions of interest in terms of model parameters, estimating the parameters from the available data, and employing appropriate inferential tools for answering the questions of interest. Before turning to issues of estimation and inference, it is important to discuss the meaning of regression coefficients and what questions can be answered through them.

DISPLAY 9.5

Model for the regression surface of flowers per plant under 12 treatment levels as a regression plane

Regression plane: $\mu\{\text{flowers} | \text{light}, \text{time}\} = \beta_0 + \beta_1 \text{light} + \beta_2 \text{time}$

$Y = \text{flowers}$
per Plant



Regression Surfaces

The multiple linear regression model with two explanatory variables, such as

$$\mu\{\text{flowers} | \text{light}, \text{time}\} = \beta_0 + \beta_1 \text{light} + \beta_2 \text{time},$$

describes the regression surface as a *plane*. The parameter β_0 is the height of the plane when both *light* and *time* equal zero; the parameter β_1 is the slope of the plane as a function of *light* for any fixed value of *time*, and the parameter β_2 is the slope of the plane as a function of *time* for any fixed value of *light*. This model is represented in Display 9.5.

When more than two explanatory variables are present, it is difficult and not generally useful to consider the geometry of regression surfaces. Instead, the regression coefficients are interpreted in terms of the effects that the selected explanatory variables have on the mean of the response when other explanatory variables are also included in the model.

Effects of Explanatory Variables

The *effect* of an explanatory variable is the change in the mean response that is associated with a one-unit increase in that variable while holding all other explanatory

variables fixed. In a regression for the meadowfoam study,

$$\begin{aligned}\text{light effect} &= \mu\{\text{flowers} | \text{light} + 1, \text{time}\} - \mu\{\text{flowers} | \text{light}, \text{time}\}, \\ \text{time effect} &= \mu\{\text{flowers} | \text{light}, \text{time} + 1\} - \mu\{\text{flowers} | \text{light}, \text{time}\}.\end{aligned}$$

In the planar model, effects are the same at all levels of the explanatory variables. The change in mean *flowers* associated with increasing *light* from 150 to 151 $\mu\text{mol}/\text{m}^2/\text{sec}$ with *time* at 24 days, for example, is the same as the change associated with increasing *light* from 600 to 601 $\mu\text{mol}/\text{m}^2/\text{sec}$ with *time* at 0 days. The precise value of the effect is found by performing the subtraction with the specified form of the model:

$$\begin{aligned}\text{light effect} &= \mu\{\text{flowers} | \text{light} + 1, \text{time}\} - \mu\{\text{flowers} | \text{light}, \text{time}\} \\ &= [\beta_0 + \beta_1(\text{light} + 1) + \beta_2\text{time}] - [\beta_0 + \beta_1\text{light} + \beta_2\text{time}] = \beta_1.\end{aligned}$$

In this model, the coefficient of one explanatory variable measures the effect of that variable at fixed values of the other.

When regression analysis is performed on the results of a randomized experiment the interpretation of an “effect” of an explanatory variable is straightforward, and causation is implied. For example, “A one-unit increase in light intensity causes the mean number of flowers to increase by β_1 .”

For observational studies, on the other hand, interpretation is more complicated. First, as usual, we cannot make causal conclusions from statistical association. Second, the *X*’s *cannot* be held fixed independently of one another as they can in a controlled experiment. We must therefore use wording like the following: “For any subpopulation of mammal species with the same body weight and litter size, a one-day increase in the species’ gestation length is associated with a β_2 gram increase in mean brain weight.” This wording requires the reader to imagine a subpopulation of mammals with fixed values of body weight and litter size, but varying gestation lengths. Whether this interpretation is useful depends on whether such subpopulations exist.

Interpretation Depends on What Other X’s Are Included

For the mammal brain weight data, furthermore, the interpretation of β_1 in the model

$$\mu\{\text{brain} | \text{gestation}\} = \beta_0 + \beta_1\text{gestation}$$

differs from the interpretation of β_1 in the model

$$\mu\{\text{brain} | \text{gestation}, \text{body}\} = \beta_0 + \beta_1\text{gestation} + \beta_2\text{body}.$$

In the first model, β_1 measures the rate of change in mean brain weight with changes in gestation length *in the population of all mammal species* . . . where body size is variable. In the second model, β_1 measures the rate of change in mean brain weight with changes in gestation length *within subpopulations of fixed body size*. If in the second model there is no effect of gestation length on mean brain weight within

subpopulations of fixed body size ($\beta_1 = 0$), and if, as expected, both brain weight and gestation length vary consistently with body size, then β_1 in the first model will not be zero. It will measure the association between brain weight and body size indirectly through their joint associations with gestation length.

9.3 SPECIALLY CONSTRUCTED EXPLANATORY VARIABLES

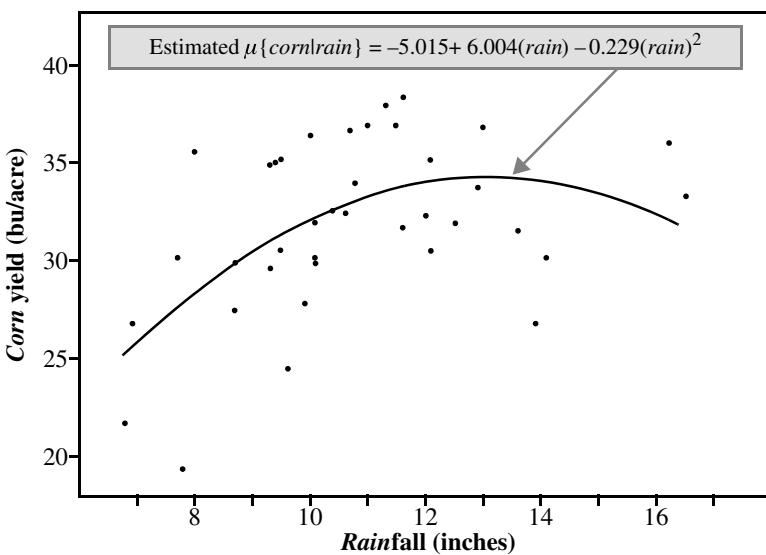
One can dramatically expand the scope of multiple linear regression by using specially constructed explanatory variables. With them, regression models can exhibit curvature, interactive effects of explanatory variables, and effects of categorical variables.

9.3.1 A Squared Term for Curvature

Display 9.6 is a scatterplot of the corn yield versus rainfall in six U.S. corn-producing states (Iowa, Nebraska, Illinois, Indiana, Missouri, and Ohio), recorded for each year from 1890 to 1927 (see also Exercise 15). A straight-line regression model is not adequate. Although increasing rainfall is associated with higher mean yields for rainfalls up to 12 inches, increasing rainfall at higher levels is associated with no change or perhaps a decrease in mean yield. In short, the rainfall effect depends on the rainfall level.

DISPLAY 9.6

Yearly corn yield versus rainfall (1890–1927) in six U.S. states



One model for incorporating curvature includes squared rainfall as an additional explanatory variable:

$$\mu\{\text{corn} \mid \text{rain}\} = \beta_0 + \beta_1 \text{rain} + \beta_2 \text{rain}^2.$$

It is not necessary that this model corresponds to any natural law (but it could). It is, however, a convenient way to incorporate curvature in the regression of *corn* on *rain*. Remember that *linear* regression means linear in β 's, not linear in X , so quadratic regression is a special case of multiple linear regression.

The model incorporates curvature by having the effect of rainfall be different at different levels of rainfall:

$$\begin{aligned}\mu\{\text{corn} \mid \text{rain} + 1\} - \mu\{\text{corn} \mid \text{rain}\} &= [\beta_0 + \beta_1(\text{rain} + 1) + \beta_2(\text{rain} + 1)^2] \\ &\quad - [\beta_0 + \beta_1 \text{rain} + \beta_2 \text{rain}^2] \\ &= \beta_1 + \beta_2[(2 \times \text{rain}) + 1].\end{aligned}$$

From the fitted model shown in Display 9.6, the effect of a unit increase in rainfall from 8 to 9 inches is estimated to be an increase in mean yield of about 2.1 bushels of corn per acre. But the effect of a unit increase in rainfall from 14 to 15 inches is estimated to be a *decrease* in mean yield of about 0.6 bu/acre.

Attempting to interpret the individual coefficients in this example is difficult and unnecessary. The statistical significance of the squared rainfall coefficient is that it highlights the inadequacy of the straight line regression model and suggests that increasing yield is associated with increasing rainfall only up to a point. In many applications squared terms are useful for incorporating slight curvature and the purpose of the analysis does not require that the coefficient of the squared term be interpreted. In specialized situations, higher-order polynomial terms may also be included as explanatory variables.

9.3.2 An Indicator Variable to Distinguish Between Two Groups

Display 9.3 shows a scatterplot of the *flower* numbers versus *light* intensity, with *time* = 0 days coded differently from *time* = 24 days. The lines on this scatterplot do not represent simple linear regression equations for *time* = 0 and *time* = 24 separately. They are the result of a multiple linear regression model that incorporates an *indicator variable* to represent the two levels of the timing variable. The model requires that the lines have equal slopes.

An *indicator variable* (or dummy variable) takes on one of two values: “1” (one) indicates that an attribute is present, and “0” (zero) indicates that the attribute is absent. The variable *early*, for example, is set equal to 1 for units where the timing was 24 days prior to PFI and is set equal to 0 for units where light intensity was not varied prior to PFI. The indicator variable *early* (along with variables to be introduced shortly) appears in Display 9.7.

DISPLAY 9.7

The meadowfoam data (first three columns) and three sets of constructed explanatory variables that are discussed in this chapter

Original variables			Timing indicators		Light level indicators						Interaction
flowers	light	time	day24	day0	L150	L300	L450	L600	L750	L900	light × day24
62.3	150	0	0	1	1	0	0	0	0	0	0
77.4	150	0	0	1	1	0	0	0	0	0	0
77.8	150	24	1	0	1	0	0	0	0	0	150
75.6	150	24	1	0	1	0	0	0	0	0	150
55.3	300	0	0	1	0	1	0	0	0	0	0
54.2	300	0	0	1	0	1	0	0	0	0	0
69.1	300	24	1	0	0	1	0	0	0	0	300
78.0	300	24	1	0	0	1	0	0	0	0	300
49.6	450	0	0	1	0	0	1	0	0	0	0
61.9	450	0	0	1	0	0	1	0	0	0	0
57.0	450	24	1	0	0	0	1	0	0	0	450
71.1	450	24	1	0	0	0	1	0	0	0	450
39.4	600	0	0	1	0	0	0	1	0	0	0
45.7	600	0	0	1	0	0	0	1	0	0	0
62.9	600	24	1	0	0	0	0	1	0	0	600
52.2	600	24	1	0	0	0	0	1	0	0	600
31.3	750	0	0	1	0	0	0	0	1	0	0
44.9	750	0	0	1	0	0	0	0	1	0	0
60.3	750	24	1	0	0	0	0	0	1	0	750
45.6	750	24	1	0	0	0	0	0	1	0	750
36.8	900	0	0	1	0	0	0	0	0	1	0
41.9	900	0	0	1	0	0	0	0	0	1	0
52.6	900	24	1	0	0	0	0	0	0	1	900
44.4	900	24	1	0	0	0	0	0	0	1	900

Consider the regression model

$$\mu\{flowers \mid light, early\} = \beta_0 + \beta_1 light + \beta_2 early.$$

If $time = 0$, then $early = 0$, and the regression line is

$$\mu\{flowers \mid light, early = 0\} = \beta_0 + \beta_1 light.$$

If $time = 24$, then $early = 1$, and the regression line is

$$\mu\{flowers \mid light, early = 1\} = \beta_0 + \beta_1 light + \beta_2.$$

This multiple regression model states that the mean number of flowers is a straight line function of light intensity for both levels of timing. The slope of both lines is β_1 ; the intercept for units with timing at PFI (*late*) is β_0 ; and the intercept for units with timing 24 days prior to PFI (*early*) is $\beta_0 + \beta_2$. Because the slopes

are the same, the model is called the *parallel lines* regression model. Furthermore, the coefficient of the indicator variable, β_2 , is the amount by which the mean number of flowers with prior timing at 24 days exceeds that with no prior timing, after accounting for the effect of light intensity differences. In Display 9.3, the lines showing (estimates for) the mean numbers of flowers versus intensity are separated by a constant vertical difference of β_2 units.

The indicator variable can be defined for either group without affecting the model. If, instead of the indicator variable *early*, the regression had been based on the indicator variable *late*, taking 1 for *time* = 0 and 0 for *time* = 24, then the resulting fit would be exactly the same; it would produce exactly the same lines on Display 9.3. The only difference would be that the intercept for units with timing prior to PFI is β_0 and the intercept for units with timing at PFI is $\beta_0 + \beta_2$. This β_2 is the negative of the β_2 in the previous version of the model.

An indicator variable may be included in a regression model just as any other explanatory variable. The coefficient of the indicator variable is the difference between the mean response for the indicated category (= 1) and the mean response for the other category (= 0), at fixed values of the other explanatory variables.

9.3.3 Sets of Indicator Variables for Categorical Explanatory Variables with More Than Two Categories

A categorical explanatory variable with more than two categories can also be incorporated into a regression model. When a categorical variable is used in regression it is called a *factor* and the individual categories are called the *levels* of the factor. If there are k levels, then $k - 1$ indicator variables are needed as explanatory variables.

Light intensity in the meadowfoam study can be viewed as a categorical variable having six levels. For use in regression, indicator variables are created for each level. Let $L150$ be equal to 1 for all experimental units that received light intensity of $150 \mu\text{mol}/\text{m}^2/\text{sec}$ and equal to 0 for all other units. Let $L300$ similarly indicate the units that received light intensity of $300 \mu\text{mol}/\text{m}^2/\text{sec}$; and so on, to $L900$, indicating the units that received light intensity of $900 \mu\text{mol}/\text{m}^2/\text{sec}$. This creates an indicator variable for every level of intensity (Display 9.7). To treat light intensity as a factor, all but one of these are included as explanatory variables in the regression model. The level whose indicator variable is not used in the set is called the *reference level* for that factor.

Intensity has six levels, so five indicator variables will represent all levels. Selecting the first level, $150 \mu\text{mol}/\text{m}^2/\text{sec}$ intensity, as the reference level, the multiple linear regression model is

$$\begin{aligned}\mu\{\text{flowers} | \text{light}, \text{day24}\} &= \beta_0 + \beta_1 L300 + \beta_2 L450 + \beta_3 L600 \\ &\quad + \beta_4 L750 + \beta_5 L900 + \beta_6 \text{day24}.\end{aligned}$$

The coefficients of the indicator variables in this model allow the mean number of flowers to vary arbitrarily with light intensity.

The questions of interest in the meadowfoam study, however, are more easily addressed with the simpler model that uses *light* as a numerical explanatory

variable, $\beta_0 + \beta_1 light + \beta_2 day24$. So, an important issue that will arise in later chapters is how best to represent a variable like light intensity—as a factor with several levels or as a numerical variable?

9.3.4 A Product Term for Interaction

Two explanatory variables are said to *interact* if the effect that one of them has on the mean response depends on the value of the other. In multiple regression, an explanatory variable for interaction can be constructed as the product of the two explanatory variables that are thought to interact.

An Interaction Model for the Meadowfoam Data

A secondary question of interest in the meadowfoam study was one of interaction: Does the effect of light intensity on mean number of flowers depend on the timing of the light regime? The product variable $light \times early$ (see Display 9.7) models an interaction between light intensity and timing. Consider the model

$$\mu\{flowers | light, early\} = \beta_0 + \beta_1 light + \beta_2 early + \beta_3(light \times early).$$

This is a way of expressing the regression of *flowers* on *light*, for different levels of timing. When *early* = 0, the regression equation is a straight-line function of intensity with intercept β_0 and slope β_1 . When *early* = 1, the mean number of flowers is also a straight line function of intensity, but with intercept $\beta_0 + \beta_2$ and slope $\beta_1 + \beta_3$. Rearranging the model as

$$\mu\{flowers | light, early\} = (\beta_0 + \beta_2 early) + (\beta_1 + \beta_3 early)light$$

shows how both the intercept and slope in the regression of *flowers* on *light* depend on the timing. The light intensity effect in the interaction model is $(\beta_1 + \beta_3 early)$; while the timing effect is $(\beta_2 + \beta_3 light)$. The effect of the light intensity depends on the timing; the effect of the timing depends on the light intensity. This representation of the regression leads directly to a graph that shows the mean of flowers as a function of light intensity, with different lines corresponding to different levels of timing, as shown in the top panel of Display 9.8.

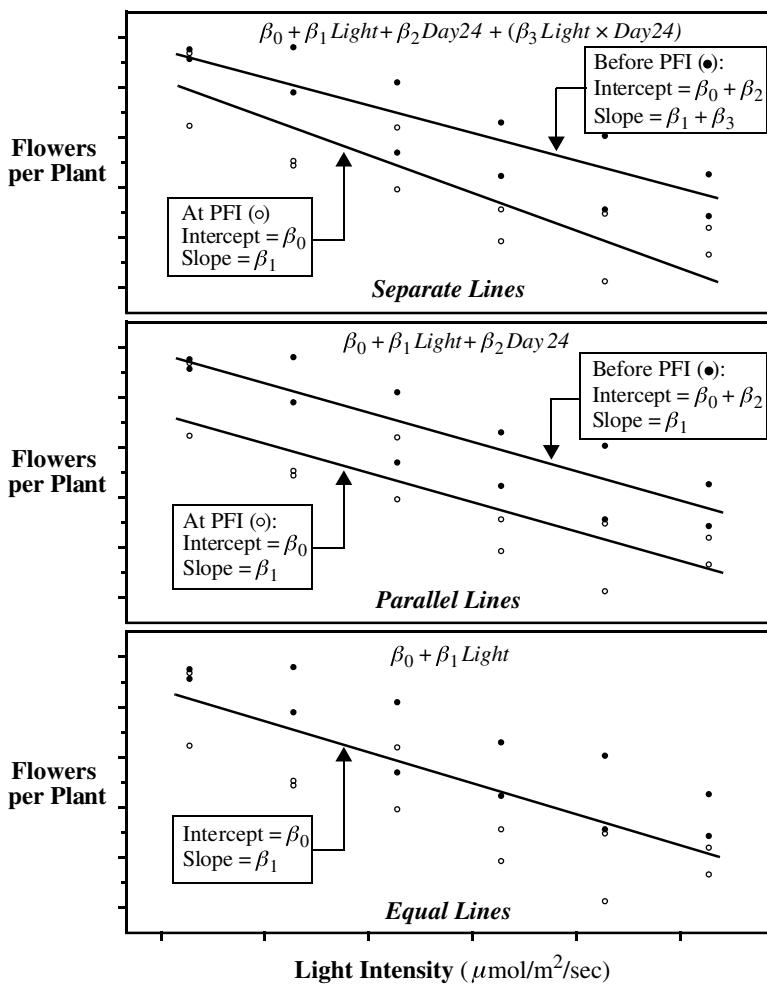
As with squared terms in models for curvature, it is often difficult to interpret individual coefficients in an interaction model. The coefficient, β_1 , of *light* has changed from being a global slope to being the slope when *time* = 0. The coefficient, β_3 , of the product term is the difference between the slope of the regression line on *light* when *time* = 24 and the slope when *time* = 0. If it is only necessary to test whether interaction is present, no lengthy interpretation is needed. The best method of communicating findings about the presence of significant interaction may be to present a table or graph of estimated means at various combinations of the interacting variables.

When to Include Interaction Terms

Interaction terms are not routinely included in regression models. Inclusion is indicated in three situations: when a question of interest pertains to interaction (as in

DISPLAY 9.8

Regression models for separate lines, parallel lines, and equal lines in two groups—meadowfoam study



the meadowfoam study); when good reason exists to suspect interaction; or when interactions are proposed as a more general model for the purpose of examining the goodness of fit of a model without interaction.

Except in special circumstances, a model including a product term for interaction between two explanatory variables should also include terms with each of the explanatory variables individually, even though their coefficients may not be significantly different from zero. Following this rule avoids the logical inconsistency of saying that the effect of X_1 depends on the level of X_2 but that there is no effect of X_1 .

Completely Separate Regression Lines for Different Levels of a Factor

Sometimes the analyst wishes to fit simple linear regressions of Y on X separately for different levels of a categorical factor. This analysis can be accomplished by repeated application of simple linear regression for each level; but it can also be accomplished with multiple regression. The multiple linear regression model has as its explanatory variables: X , the $k - 1$ indicators to distinguish the k levels, and all products of the indicators with X .

The multiple regression approach has advantages. Questions about similarities among the separate regressions are expressible as hypotheses about specific parameters in the multiple regression model: Are the slopes all equal? Are the intercepts equal? Furthermore, the multiple regression approach conveniently provides a single, combined estimate of variance, which is equivalent to the pooled estimate from the separate simple regressions.

Display 9.8 depicts three models for the meadowfoam study, each specifying that the mean number of flowers is a straight line function of light intensity. The separate regression lines model in the top panel is the most general of the three. If the coefficient of the interaction term, β_3 , is zero, then the separate regression lines model reduces to the parallel regression lines model, which indicates no interaction of *light* and *time*. Similarly, if β_2 is zero in the parallel regression lines model, then it reduces to the equal lines model.

9.3.5 A Shorthand Notation for Model Description

The notation for regression models is sometimes unnecessarily bulky when there are a large number of indicator variables or interaction terms. An abbreviation for specifying a categorical explanatory variable is to list the categorical variable name in uppercase letters to represent the entire set of indicator variables used to model it. For example, $\mu\{\text{flowers}, \text{LIGHT}, \text{early}\}$ indicates that light intensity is a factor, whose effects are to be modeled by indicator variables.

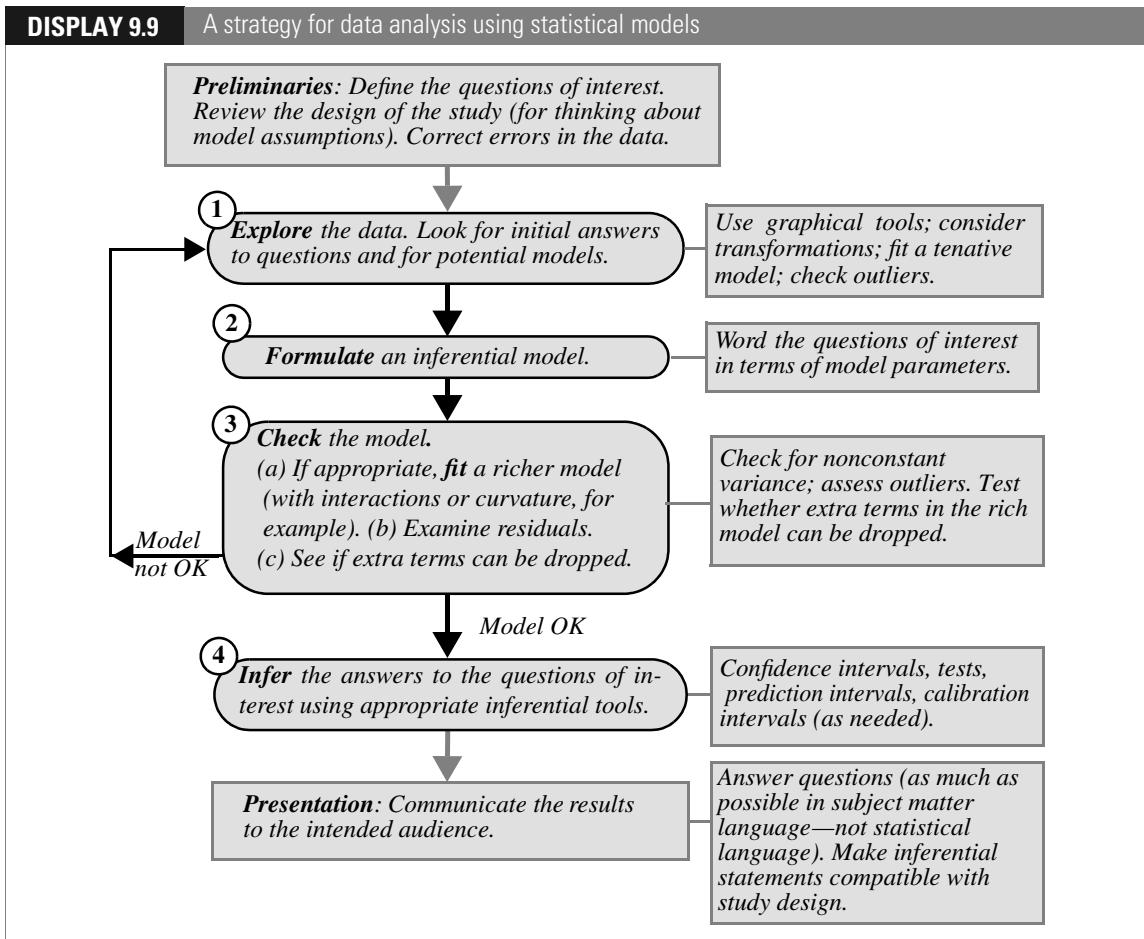
Additional savings are achieved by omitting the parameters when specifying the terms included in the model. For example,

$$\mu\{\text{flowers} \mid \text{light}, \text{TIME}\} = \text{light} + \text{TIME}$$

describes the parallel lines model with *light* treated as a numerical explanatory variable and *TIME* as a factor. Similarly,

$$\mu\{\text{flowers} \mid \text{light}, \text{TIME}\} = \text{light} + \text{TIME} + (\text{light} \times \text{TIME})$$

specifies the separate lines model with the effects of the numerical explanatory variable *light*, the factor *TIME*, and the interaction terms formed as the products of *light* with the indicator variable(s) for the factor *TIME*. Extensions of this notation conveniently describe models with any number of numerical and categorical variables.



9.4 A STRATEGY FOR DATA ANALYSIS

Display 9.9 shows a general scheme for using statistical models to analyze data. Data analysis involves more than inserting data into a computer and pressing the right button. Substantial exploratory analysis may be required at the outset, and several dead ends may be encountered in the pursuit of suitable models.

The strategy for data analysis centers on the development of an *inferential model* where answers to the questions of interest can be found by drawing inferences about key parameters. The choice of an inferential model may be guided by the graphical displays. Further checking on model adequacy is accomplished by residual analysis and by informal testing of terms in a richer model that contains additional parameters representing possible curvature, interactions, or more complex features.

The adequacy of a model may be clear from the graphical displays and residual analysis alone. When it is not, informal testing is an important part of finding an

adequate model. In particular, if replicates of the response occur at each combination of explanatory variables (as in the meadowfoam study), it is always possible to compare the model that is convenient for analysis to the separate-means model (the one with a separate response mean at each combination of the explanatory variables). Moreover, suspicions of nonlinearities and interactions can be investigated by testing the appropriate additional terms that model these inadequacies. If a particular, convenient inferential model is found to be inappropriate, the analyst has two options: renew the search for a better inferential model, or offer a more complicated explanation of the study's conclusions in terms of the richer model.

Ensuing chapters provide examples of this strategy. Readers should refer to Display 9.9 often as the analysis patterns develop. The remainder of this chapter deals with graphical exploratory techniques (for steps 1 and 2 of this strategy). The next chapter details the estimation of parameters and the main inferential tools (for steps 3 and 4). Chapter 11 addresses new issues and techniques for model assessment (for step 3). Chapter 12 discusses the difficulties with a set containing too many explanatory variables and some computer-assisted techniques for reducing the set to a reasonable number (for steps 1 and 2).

9.5 GRAPHICAL METHODS FOR DATA EXPLORATION AND PRESENTATION

9.5.1 A Matrix of Pairwise Scatterplots

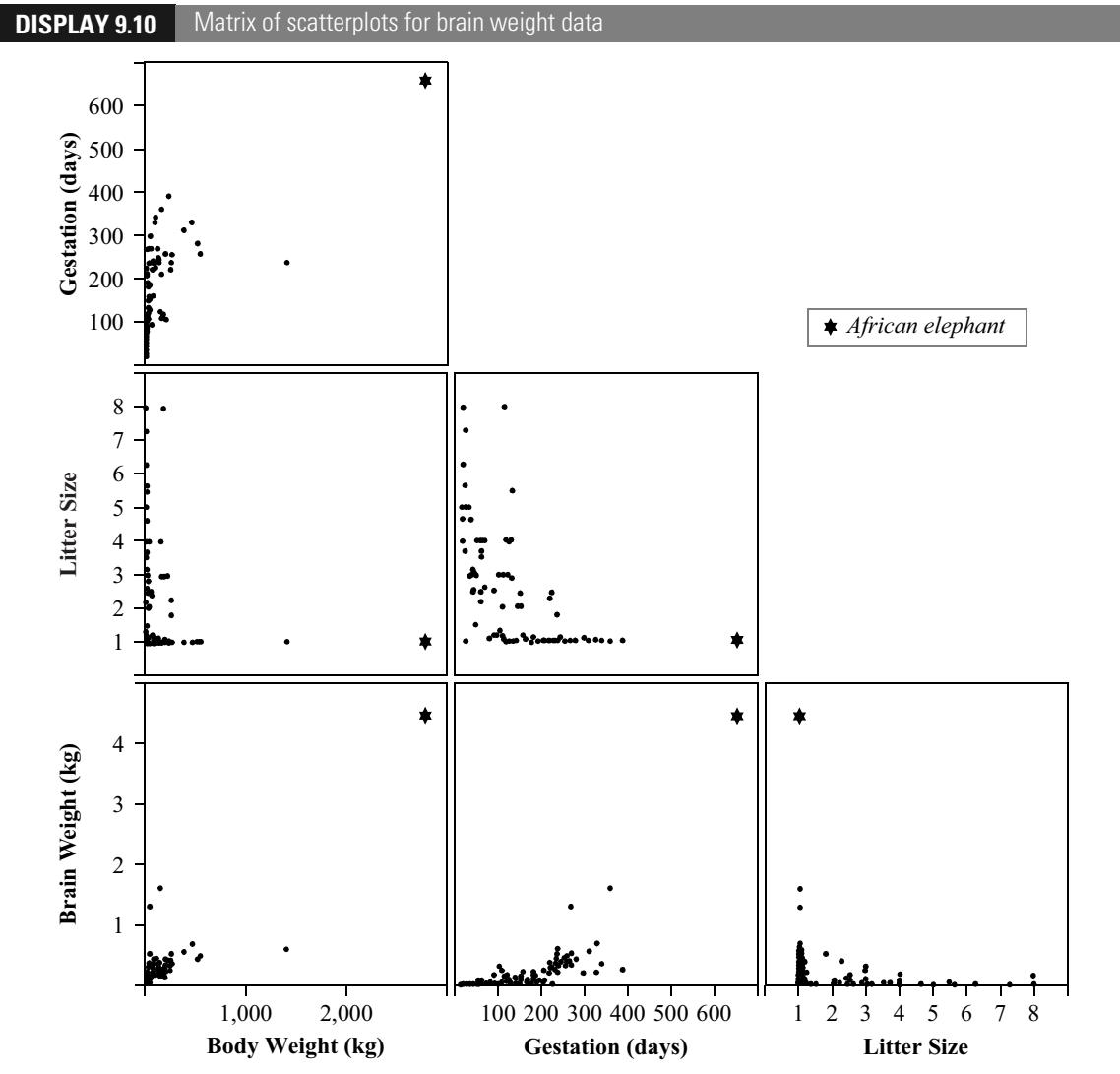
Individual scatterplots of the response variable versus each of the explanatory variables are usually helpful. Although the observed relationships in these plots will not necessarily indicate the effects of interest in a *multiple* regression model, the plots are still useful for observing the marginal (one at a time) relationships, drawing attention to interesting points, and suggesting the need for transformations.

A *matrix of scatterplots* (or *draftsman plot*) is a consolidation of all possible pairwise scatterplots from a set of variables, as shown for the mammal brain weight data in Display 9.10. The bottom row shows the response variable versus each of the explanatory variables and the higher rows show the scatterplots of the explanatory variables versus each other. There is a lot to look at in such a display. Typically, one would investigate the scatterplots of the response versus each of the explanatory variables (the bottom row) first.

Different statistical computer packages have different display formats. Instead of the lower left triangular arrangement shown in Display 9.10, some packages display a corresponding upper right triangle (in addition to the lower left one) that contains the same plots but with axes reversed. This does not contain any more information, although it may provide a different visual perspective. The resolution with which the plots can be drawn may also be an issue. It becomes increasingly difficult to read a matrix of scatterplots as the number of variables is increased. For data sets with many potential explanatory variables, it may be necessary to select subsets for display.

DISPLAY 9.10

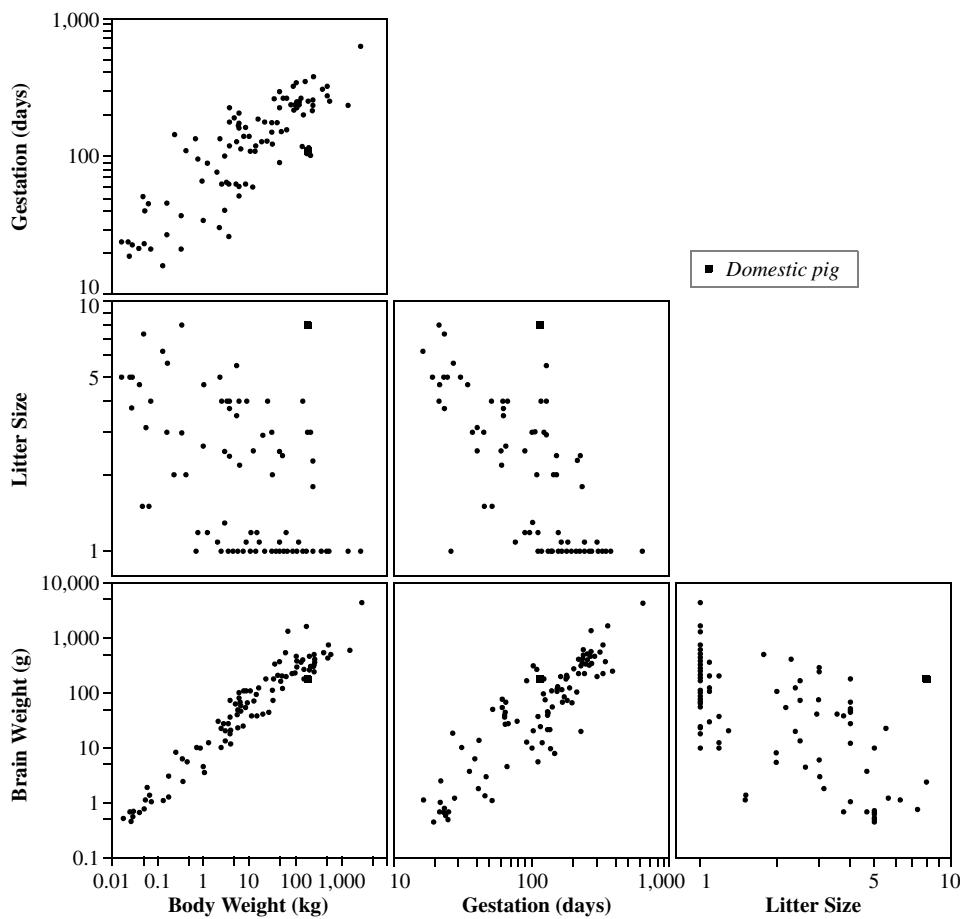
Matrix of scatterplots for brain weight data

**Notes About the Scatterplots for the Brain Weight Data**

Consider the bottom row of Display 9.10 first. The scatterplot of brain weight versus body weight is not very helpful since most of the data points are clustered on top of each other in the bottom left-hand corner. The upper bounds for the X and Y axes are determined largely by one point (the African elephant). This scatterplot indicates what should have been evident from the start—that the mammals in this list differ in size by orders of magnitude. The display should be redrawn after brain weight and body weight are transformed to their logarithms.

DISPLAY 9.11

Matrix of scatterplots for brain weight data, after log transformations



Gestation values also appear to be quite skewed. A trial plot using the log of this variable shows that it too should be transformed. Display 9.11 shows a revised version of the matrix of scatterplots where all variables have been placed on their natural logarithmic scales.

The bottom row of Display 9.11 shows less overlap and a more distinct pattern. Notice the pronounced relationship between log brain weight and each of the explanatory variables. Notice also, from the upper plots in the first column, that gestation and litter size are also related to body weight. Therefore the question of whether there is an association between gestation and brain weight after accounting for the effect of body weight, and the question of whether there is an association between litter size and brain weight after accounting for the effect of body weight, are not resolved by these plots. Nevertheless, they suggest the next course of action,

which is to fit a regression model for log brain weight on log body weight, log gestation, and log litter size.

In the scatterplot of log brain weight versus log litter size, there is one data point that stands out—a mammal with litter size eight whose brain size is quite a bit larger than those for other mammals whose litter size is eight. The analyst should determine which mammal this is. It may be influential in the analysis, and it may provide useful information or additional questions for research. Modern statistical computer packages with interactive labeling features make the identification of such points on a graph fairly easy.

9.5.2 Coded Scatterplots

Display 9.3 is an example of a *coded scatterplot*—a scatterplot with different symbols or letters used as plotting marks to distinguish two or more groups. In the example, the groups correspond to different levels of timing. This is a great way of observing the joint effects of one numerical and one categorical explanatory variable. For observing the joint effects of two numerical explanatory variables, it is often very helpful to plot the response versus one of them, and to use a plotting code to represent groups defined by ranges of the other (such as educational level less than 8 years, between 9 and 12 years, or greater than 12 years).

9.5.3 Jittered Scatterplots

The scatterplot of log brain weight versus log litter size (Display 9.11) contains many points that overlap because many species of mammals have the same litter size. This makes it difficult to judge relative frequency. When a variable has only a few distinct values it may be possible to improve the visual information of the plot by *jittering* that variable—adding small computer-generated random numbers to it before plotting. In Display 9.12, log brain weight is plotted against a jittered version of litter size. A jittered variable is only used in the scatterplot; subsequent numerical analyses must be based on the unjittered version. The choice as to the size of the random numbers to add in order to make each point visually distinct requires trial and error.

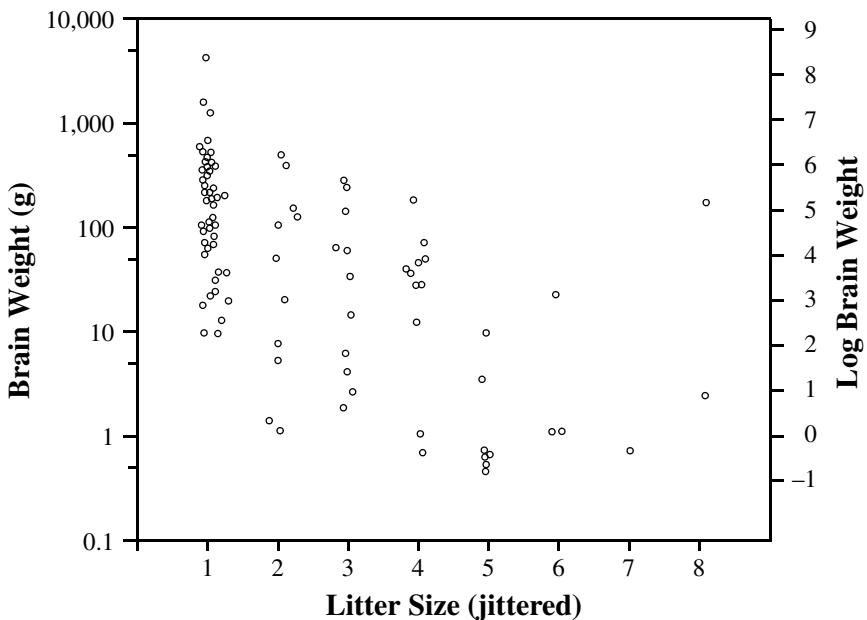
9.5.4 Trellis Graphs

In Section 9.2 it was noted that β_2 in the model $\mu\{Y|X_1, X_2\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ refers to the change in the mean of Y associated with changes in X_2 , for subpopulations of fixed X_1 . A useful graphical procedure for exploring this relationship is a trellis graph. A simple form is shown in Display 9.13, in which the response variable, log brain weight, is plotted against an explanatory variable of interest—log of gestation length—separately for species in the lowest 25% of body weights, the next lowest, and so on.

While the data in the four panels are not subpopulations with fixed body weights, they are subpopulations with similar values. This display suggests a positive relationship exists between brain weight and gestation length even accounting for different body weights. Furthermore, the slope in this relationship appears to

DISPLAY 9.12

Jittered scatterplot: log brain weight versus litter size (jittered)



be the same in the different panels of body weight; that is, there is no evidence of an interactive effect of gestation and body weight on brain weight.

9.6 RELATED ISSUES

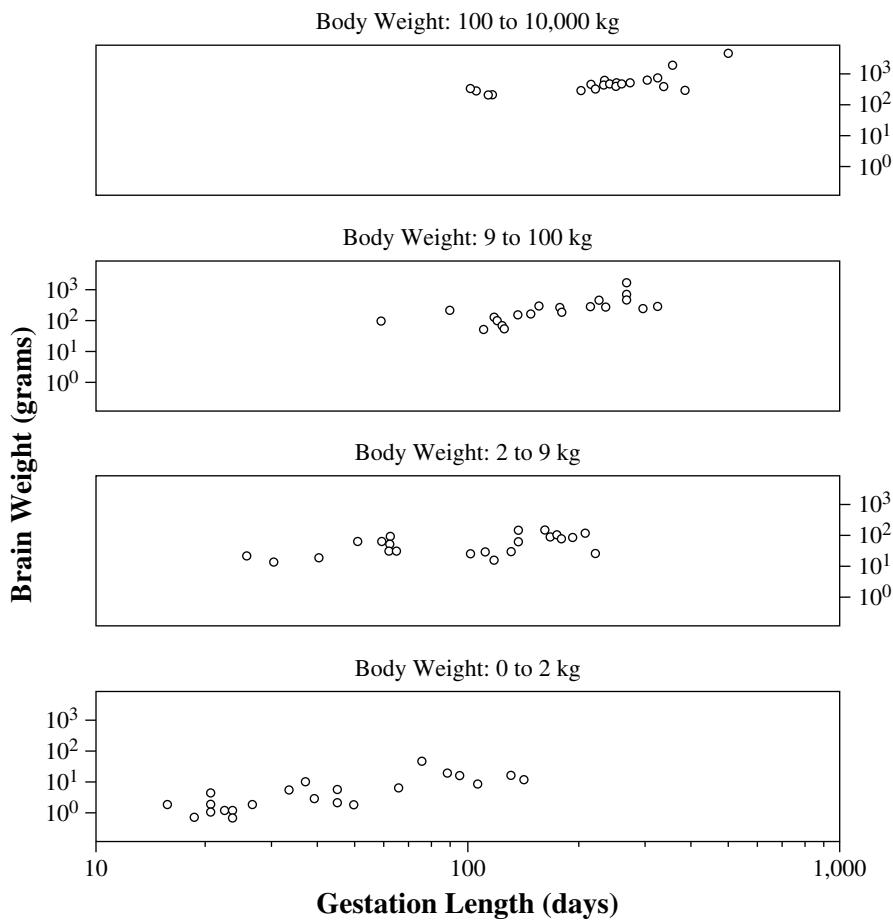
9.6.1 Computer Output

Regression output from statistical computer packages can be intimidating for the simple reason that it tends to include everything that the user could possibly want and more. One important component, which all packages display, is the list of estimates and their standard errors. Display 9.14 and Display 9.15 show the least squares estimates of the coefficients in specified models for the two examples of this chapter, their standard errors, and two-sided p -values from t -tests for the hypotheses that each β is zero in the specified model.

Although the least squares estimates and the inferential tools are discussed further in the next chapter, it is not difficult to see which questions are addressed by the p -values. The p -value of 0.9096 for the interaction term from Display 9.14 indicates that zero is a plausible value for the coefficient of the interaction explanatory variable; so there is no evidence of an interaction. The p -value of 0.0089 for the coefficient of *litter* indicates strong evidence that the litter size is associated with brain weight, even after accounting for body weight and gestation.

DISPLAY 9.13

Trellis graph showing scatterplots of log brain weight versus log gestation length, separately for species in each of four categories of body weight



9.6.2 Factorial Treatment Arrangement

The meadowfoam study investigated two treatment factors applied jointly to each experimental unit: the light intensity (at one of six levels) and the timing (at one of two levels). All 12 combinations of these two treatment factors were applied to some experimental units. The term *factorial treatment arrangement* describes the formation of experimental treatments from all combinations of the levels of two or more distinct, individual treatment factors. This study was a 6×2 factorial arrangement, meaning all six levels of one factor appear in combination with both levels of the other.

Three benefits of testing multiple factors all at once rather than through one-at-a-time experiments are the following: Interactive effects can be explored; there is

DISPLAY 9.14

Estimates of regression coefficients in the multiple regression of *flowers* on *light*, *early*, and *light* × *early*—the meadowfoam study

Variable	Coefficient	Standard error	t-statistic	p-value
Constant	71.6233	4.3433	16.4905	<0.0001
<i>light</i>	-0.0411	0.0074	5.5247	<0.0001
<i>early</i>	11.5233	6.1424	1.8760	0.0753
<i>light</i> × <i>early</i>	0.0012	0.0105	0.1150	0.9096

DISPLAY 9.15

Estimates of regression coefficients in the multiple regression of log brain weight on log body weight, log gestation, and log litter size—brain weight data

Variable	Coefficient	Standard error	t-statistic	p-value
Constant	0.8548	0.6617	1.2919	0.1996
<i>lbody</i>	0.5751	0.0326	17.6468	<0.0001
<i>lgest</i>	0.4179	0.1408	2.9687	0.0038
<i>llitter</i>	-0.3101	0.1159	2.6747	0.0089

more efficient use of the available experimental units with the multifactor arrangement (resulting in smaller standard errors for parameters of interest); and the results for the multifactor experiment are more general, since each treatment is investigated at several levels of the other treatment.

9.7 SUMMARY

Multiple regression analysis refers to a large set of tools associated with developing and using regression models for answering questions of interest. Multiple regression models describe the mean of a single response variable as a function of several explanatory variables. Transformations, indicator variables for grouped data (factors), squared explanatory variables for curvature, and product terms for interaction greatly enhance the usefulness of this model.

Substantial exploratory analysis is recommended for gaining initial insight into what the data have to say in answer to the questions of interest and for suggesting possible regression models for answering them more formally. Some standard graphical procedures are presented here. The data analyst should be prepared to be creative in using the available computer tools to best display the data, while keeping in mind the questions of interest and the statistical tools that might be useful for answering them.

Brain Weight Study

Is brain weight associated with gestation period and/or litter size after accounting for the effect of body weight? This is exactly the type of question for which multiple

regression is useful. The regression coefficient of gestation in the regression of brain weight on body weight and gestation describes the effect of gestation for species of roughly the same body weight. With multiple linear regression, the coefficients can be estimated from all the animals without a need for grouping into subsets of similar body weight. Initial scatterplots (or initial inspection of the data) indicate that the regression model should be formed after transforming all the variables to their logarithms.

Meadowfoam Study

A starting point for the analysis is the coded scatterplot of number of flowers per plant versus light intensity, with different codes to represent the two levels of the timing factor (Display 9.3). The plot suggests that the mean number of flowers per plant decreases with increasing light intensity, that the rate of decrease does not depend on timing, and that (for any light intensity value) a larger mean number of flowers is associated with the “before PFI” level of the timing factor. Using an indicator variable for one of the timing levels permits fitting the parallel regression lines model. Further inclusion of the interaction of timing and intensity produces a model that fits separate regression lines for each level of timing. This model permits a check on whether the regression lines are indeed parallel.

9.8 EXERCISES

Conceptual Exercises

1. **Meadowfoam.** (a) Write down a multiple regression model with parallel regression lines of *flowers* on *light* for the two separate levels of *time* (using an indicator variable). (b) Add a term to the model in (a) so that the regression lines are not parallel.
2. **Meadowfoam.** A model (without interaction) for the mean *flowers* is estimated to be $71.3058 - 0.0405\text{light} + 12.1583\text{early}$. For a fixed level of timing, what is the estimated difference between the mean *flowers* at 600 and 300 $\mu\text{mol}/\text{m}^2/\text{sec}$ of *light* intensity?
3. **Meadowfoam.** (a) Why were the numbers of flowers from 10 plants averaged to make a response, rather than representing them as 10 different responses? (b) What assumption is assisted by averaging the numbers from the 10 plants?
4. **Mammal Brain Weights.** The three-toed sloth has a gestation period of 165 days. The Indian fruit bat has a gestation period of 145 days. From Display 9.14 the estimated model for the mean of log brain weight is $0.8548 + 0.5751\text{log body} + 0.4179\text{log gest} - 0.3101\text{litter}$. Since *log gest* for the sloth is 0.1292 more than *log gest* for the fruit bat, does this imply that an estimate of the mean log brain weight for the sloth is $(0.4179)(0.1292)$ more than the mean log brain weight for the bat (i.e., the median is 5.5% higher)? Why? Why not?
5. **Insulating Fluid** (Section 8.1.2). Would it be possible to test for lack of fit to the straight line model for the regression of log breakdown time on voltage by including a voltage-squared term in the model, and testing whether the coefficient of the squared term is zero?
6. **Island Area and Species.** For the island area and number of species data in Section 8.1.1, would it be possible to test for lack of fit to the straight line model for the regression of log number

of species on log island area by including the square of log area in the model and testing whether its coefficient is zero?

7. Which of the following regression models are *linear*?
 - (a) $\mu\{Y|X\} = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$
 - (b) $\mu\{Y|X\} = \beta_0 + \beta_1 10^X$
 - (c) $\mu\{Y|X\} = (\beta_0 + \beta_1 X)/(\beta_0 + \beta_2 X)$
 - (d) $\mu\{Y|X\} = \beta_0 \exp(\beta_1 X)$.
8. Describe what σ measures in the meadowfoam problem and in the brain weight problem.
9. **Pollen Removal.** Reconsider the data on proportion of pollen removed and duration of visit to the flower for bumblebee queens and honeybee workers, in Exercise 3.28. (a) Write down a model that describes the mean proportion of pollen removed as a straight-line function of duration of visit, with separate intercepts and separate slopes for bumblebee queens and honeybee workers. (b) How would you test whether the effect of duration of visit on proportion removed is the same for queens as for workers?
10. **Breast Milk and IQ.** In a study, intelligence quotient (IQ) test scores were obtained for 300 8-year-old children who had been part of a study of premature babies in the early 1980s. Because they were premature, all the babies were fed milk by a tube. Some of them received breast milk entirely, some received a prepared formula entirely, and some received some combination of breast milk and formula. The proportion of breast milk in the diet depended on whether the mother elected to provide breast milk and to what extent she was successful in expressing any, or enough, for the baby's diet. The researchers reported the results of the regression of the response variable—IQ at age 8—on social class (ordered from 1, the highest, to 5), mother's education (ordered from 1, the lowest, to 5), an indicator variable taking the value 1 if the child was female and 0 if male, the number of days of ventilation of the baby after birth, and an indicator variable taking the value 1 if there was any breast milk in the baby's diet and 0 if there was none. The estimates are reported in Display 9.16 along with the p -values for the tests that each coefficient is zero. (Data from Lucas et al., "Breast Milk and Subsequent Intelligence Quotient in Children Born Preterm," *Lancet* 339 (1992): 261–64).

DISPLAY 9.16

Breast milk and intelligence data

Explanatory variable	Estimated coefficient	<i>p</i> -value
Social class	−3.5	0.0004
Mother's education	2.0	0.01
Female indicator	4.2	0.01
Days of ventilation	−2.6	0.02
Breast milk indicator	8.3	<0.0001

- (a) After accounting for the effects of social class, mother's education, whether the child was a female, and days after birth of ventilation, how much higher is the estimated mean IQ for those children who received breast milk than for those who did not?
- (b) Is it appropriate to use the variables "Social class" and "Mother's education" in the regression even though in both instances the numbers 1 to 5 do not correspond to anything real but are merely ordered categories?
- (c) Does it seem appropriate for the authors to simply report < 0.0001 for the p -value of the breast milk coefficient rather than the actual p -value?

- (d) Previous studies on breast milk and intelligence could not separate out the effects of breast milk and the act of breast feeding (the bonding from which might encourage intellectual development of the child). How is the important confounding variable of whether a child is breast fed dealt with in this study?
- (e) Why is it important to have social class and mother's education as explanatory variables?
- (f) In a subsidiary analysis the researchers fit the same regression model as above except with the indicator variable for whether the child received breast milk replaced by the percentage of breast milk in the diet (between 0 and 100%). The coefficient of that variable turned out to be 0.09. (i) From this model, how much larger is the estimated mean IQ for children who received 100% breast milk than for those who received 50% breast milk, after accounting for the other explanatory variables? (ii) What is the importance of the percentage of breast milk variable in dealing with confounding variables?

11. Glasgow Graveyards. Do persons of higher socioeconomic standing tend to live longer? This was addressed by George Davey Smith and colleagues through the relationship of the heights of commemoration obelisks and the life lengths of the corresponding grave site occupants. In burial grounds in Glasgow a certain design of obelisk is quite prevalent, but the heights vary greatly. Since the height would influence the cost of the obelisk, it is reasonable to believe that height is related to socioeconomic status. The researchers recorded obelisk height, year of death, age at death, and gender for 1,349 individuals who died prior to 1921. Although they were interested in the relationship between mean life length and obelisk height, it is important that they included year of construction as an explanatory variable since life lengths tended to increase over the years represented (1801 to 1920). For males, they fit the regression of life length on obelisk height (in meters) and year of obelisk construction and found the coefficient of obelisk height to be 1.93. For females they fit the same regression and found the coefficient of obelisk height to be 2.92. (Data from Smith et al., "Socioeconomic Differentials in Mortality: Evidence from Glasgow Graveyards," *British Medical Journal* 305 (1992): 1557–60.)

- (a) After accounting for year of obelisk construction, each extra meter in obelisk height is associated with Z extra years in mean lifetime. What is the estimated Z for males? What is the estimated Z for females?
- (b) Since the coefficients differ significantly from zero, would it be wise for an individual to build an extremely tall obelisk, to ensure a long life time?
- (c) The data were collected from eight different graveyards in Glasgow. Since there is a potential blocking effect due to the different graveyards, it might be appropriate to include a graveyard effect in the model. How can this be done?

Computational Exercises

12. Mammal Brain Weights. (a) Draw a matrix of scatterplots for the mammal brain weight data (Display 9.4) with all variables transformed to their logarithms (to reproduce Display 9.11). (b) Fit the multiple linear regression of log brain weight on log body weight, log gestation, and log litter size, to confirm the estimates in Display 9.15. (c) Draw a matrix of scatterplots as in (a) but with litter size on its natural scale (untransformed). Does the relationship between log brain weight and litter size appear to be any better or any worse (more like a straight line) than the relationship between log brain weight and log litter size?

13. Meat Processing. One way to check on the adequacy of a linear regression is to try to include an X -squared term in the model to see if there is significant curvature. Use this technique on the meat processing data of Section 7.1.2. (a) Fit the multiple regression of pH on hour and hour-squared. Is the coefficient of hour-squared significantly different from zero? What is the p -value? (b) Fit the

multiple regression of pH on log(hour) and the square of log(hour). Is the coefficient of the squared-term significantly different from zero? What is the *p*-value? (c) Does this exercise suggest a potential way of checking the appropriateness of taking the logarithm of *X* or of leaving it untransformed?

14. Pace of Life and Heart Disease. Some believe that individuals with a constant sense of time urgency (often called type-A behavior) are more susceptible to heart disease than are more relaxed individuals. Although most studies of this issue have focused on individuals, some psychologists have investigated geographical areas. They considered the relationship of city-wide heart disease rates and general measures of the pace of life in the city.

For each region of the United States (Northeast, Midwest, South, and West) they selected three large metropolitan areas, three medium-size cities, and three smaller cities. In each city they measured three indicators of the pace of life. The variable *walk* is the walking speed of pedestrians over a distance of 60 feet during business hours on a clear summer day along a main downtown street. *Bank* is the average time a sample of bank clerks takes to make change for two \$20 bills or to give \$20 bills for change. The variable *talk* was obtained by recording responses of postal clerks explaining the difference between regular, certified, and insured mail and by dividing the total number of syllables by the time of their response. The researchers also obtained the age-adjusted death rates from ischemic heart disease (a decreased flow of blood to the heart) for each city (*heart*). The data in Display 9.17 were read from a graph in the published paper. (Data from R. V. Levine, "The Pace of Life," *American Scientist* 78 (1990): 450–9.) The variables have been standardized, so there are no units of measurement involved.

DISPLAY 9.17

First five rows of the pace-of-life data set with bank clerk speed, pedestrian walking speed, postal clerk talking speed, and age-adjusted death rates due to heart disease, in 36 cities

City	Bank	Walk	Talk	Heart
Atlanta, GA	25	27	27	19
Bakersfield, CA	29	18	25	11
Boston, MA	31	28	24	24
Buffalo, NY	30	23	23	29
Canton, OH	28	20	18	19

- (a) Draw a matrix of scatterplots of the four variables. Construct it so that the bottom row of plots all have *heart* on the vertical axis. If you do not have this facility, draw scatterplots of *heart* versus each of the other variables individually.
- (b) Obtain the least squares fit to the linear regression of *heart* on *bank*, *walk*, and *talk*.
- (c) Plot the residuals versus the fitted values. Is there evidence that the variance of the residuals increases with increasing fitted values or that there are any outliers?
- (d) Report a summary of the least squares fit. Write down the estimated equation with standard errors below each estimated coefficient.

15. Rainfall and Corn Yield. The data on corn yields and rainfall, discussed in Section 9.3.1, appear in Display 9.18. (Data from M. Ezekiel and K. A. Fox, *Methods of Correlation and Regression Analysis*, New York: John Wiley & Sons, 1959; originally from E. G. Misner, "Studies of the Relationship of Weather to the Production and Price of Farm Products, I. Corn" [mimeographed publication, Cornell University, March 1928].)

- (a) Plot corn yield versus rainfall.
- (b) Fit the multiple regression of corn yield on *rain* and *rain*².

DISPLAY 9.18

Partial listing of a data set with average corn yield in a year (bushels) and total rainfall (inches) in six U.S. states (1890–1927).

Year	Yield	Rainfall
1890	24.5	9.6
1891	33.7	12.9
1892	27.9	9.9
1893	27.5	8.7
1894	21.7	6.8
...		
1927	32.6	10.4

- (c) Plot the residuals versus year. Is there any pattern evident in this plot? What does it mean? (Anything to do, possibly, with advances in technology?)
- (d) Fit the multiple regression of corn yield on $rain$, $rain^2$, and $year$. Write the estimated model and report standard errors, in parentheses, below estimated coefficients. How do the coefficients of $rain$ and $rain^2$ differ from those in the estimated model in (b)? How does the estimate of σ differ? (larger or smaller?) How do the standard errors of the coefficients differ? (larger or smaller?) Describe the effect of an increase of one inch of rainfall on the mean yield over the range of rainfalls and years.
- (e) Fit the multiple regression of corn yield on $rain$, $rain^2$, $year$, and $year \times rain$. Is the coefficient of the interaction term significantly different from zero? Could this term be used to say something about technological improvements regarding irrigation?
- 16. Pollen Removal.** The data in Exercise 3.27 are the proportions of pollen removed and the duration of visits on a flower for 35 bumblebee queens and 12 honeybee workers. It is of interest to understand the relationship between the proportion removed and duration and the relative pollen removal efficiency of queens and workers. (a) Draw a coded scatterplot of proportion of pollen removed versus duration of visit; use different symbols or letters as the plotting codes for queens and workers. Does it appear that the relationship between proportion removed and duration is a straight line? (b) The logit transformation is often useful for proportions between 0 and 1. If p is the proportion then the logit is $\log[p/(1 - p)]$. This is the log of the ratio of the amount of pollen removed to the amount not removed. Draw a coded scatterplot of the logit versus duration. (c) Draw a coded scatterplot of the logit versus log duration. From the three plots, which transformations appear to be worthy of pursuing with a regression model? (d) Fit the multiple linear regression of the logit of the proportion of pollen removed on (i) log duration, (ii) an indicator variable for whether the bee is a queen or a worker, and (iii) a product term for the interaction of the first two explanatory variables. By examining the p -value of the interaction term, determine whether there is any evidence that the proportion of pollen depends on duration differently for queens than for workers. (e) Refit the multiple regression but without the interaction term. Is there evidence that, after accounting for the amount of time on the flower, queens tend to remove a smaller proportion of pollen than workers? Why is the p -value for the significance of the indicator variable so different in this model than in the one with the interaction term?
- 17. Crab Claw and Force.** Using the crab data from Exercise 7.22, (a) draw a scatterplot of claw closing force versus propodus height (both on a log scale), with different plotting symbols to distinguish the three different crab species, and (b) fit the multiple regression of log force on log height and species (as a factor). Provide the estimated model including standard errors of estimated

DISPLAY 9.19

First five rows of a data set with average wing sizes of male and female flies, on logarithmic scale, with standard errors; and average basal length to wing size ratios of the females, at 11 locations in North America and 10 locations in Europe

Continent	Latitude (N)	Wing size ($10^3 \times \log \text{ mm}$)				Basal length to wing size (females)	
		Females	SE	Males	SE	Ratio (av)	SE
NA	35.5	901	2.5	797	3.8	0.831	0.010
NA	37.0	896	3.5	806	3.0	0.834	0.014
NA	38.6	906	3.0	812	3.2	0.836	0.012
NA	40.7	907	3.5	807	3.2	0.833	0.013
NA	40.9	898	3.6	818	2.7	0.830	0.012

regression coefficients. (See Exercises 10.9 and 10.10 for analyses that explore a more sophisticated model for these data.)

18. Speed of Evolution. How fast can evolution occur in nature? Are evolutionary trajectories predictable or idiosyncratic? To answer these questions R. B. Huey et al. (“Rapid Evolution of a Geographic Cline in Size in an Introduced Fly,” *Science* 287 (2000): 308–9) studied the development of a fly—*Drosophila subobscura*—that had accidentally been introduced from the Old World into North America (NA) around 1980. In Europe (EU), characteristics of the flies’ wings follow a “cline”—a steady change with latitude. One decade after introduction, the NA population had spread throughout the continent, but no such cline could be found. After two decades, Huey and his team collected flies from 11 locations in western NA and native flies from 10 locations in EU at latitudes ranging from 35–55 degrees N. They maintained all samples in uniform conditions through several generations to isolate genetic differences from environmental differences. Then they measured about 20 adults from each group. Display 9.19 shows average wing size in millimeters on a logarithmic scale, and average ratios of basal lengths to wing size.

- (a) Construct a scatterplot of average wing size against latitude, in which the four groups defined by continent and sex are coded differently. Do these suggest that the wing sizes of the NA flies have evolved toward the same cline as in EU?
- (b) Construct a multiple linear regression model with wing size as the response, with latitude as a linear explanatory variable, and with indicator variables to distinguish the sexes and continents. As there are four groups, you will want to have three indicator variables: the continent indicator, the sex indicator, and the product of the two. Construct the model in such a way that one parameter measures the difference between the slopes of the wing size versus latitude regressions of NA and EU for males, one measures the difference between the NA–EU slope difference for females and that for males, one measures the difference between the intercepts of the regressions of NA and EU for males, and one measures the difference between the NA–EU intercepts’ difference for females and that for males.

19. Depression and Education. Has homework got you depressed? It could be worse. Depression, like other illnesses, is more prevalent among adults with less education than you have.

R. A. Miech and M. J. Shanahan investigated the association of depression with age and education, based on a 1990 nationwide (U.S.) telephone survey of 2,031 adults aged 18 to 90. Of particular interest was their finding that the association of depression with education strengthens with increasing age—a phenomenon they called the “divergence hypothesis.”

DISPLAY 9.20

Kentucky Derby winners, 1896–2011. *Starters* is the number of horses that started the race; *NetToWinner* is the net winnings in dollars; *Time* is the winning time in seconds; *Speed* is the winning average speed, in miles per hour; *Track* is a categorical variable describing track conditions, with seven categories: Fast, Good, Dusty, Slow, Heavy, Muddy, and Sloppy; *Conditions* is a two-category version of *Track*, with categories Fast (which combines categories Fast and Good from *Track*) and Slow (which combines all other categories of *Track*); partial listing.

Year	Winner	Starters	NetToWinner	Time	Speed	Track	Conditions
1896	Ben Brush	8	4,850	127.75	35.23	Dusty	Fast
1897	Typhoon II	6	4,850	132.50	33.96	Heavy	Slow
1898	Plaudit	4	4,850	129.00	34.88	Good	Fast
1899	Manuel	5	4,850	132.00	34.09	Fast	Fast
1900	Lieut. Gibson	7	4,850	126.25	35.64	Fast	Fast
...							
2011	Animal Kingdom	20	2,000,000	122.04	36.87	Fast	Fast

They constructed a depression score from responses to several related questions. Education was categorized as (i) college degree, (ii) high school degree plus some college, or (iii) high school degree only. (See “Socioeconomic Status and Depression over the Life Course,” *Journal of Health and Social Behaviour* 41(2) (June, 2000): 162–74.)

- (a) Construct a multiple linear regression model in which the mean depression score changes linearly with age in all three education categories, with possibly unequal slopes and intercepts. Identify a single parameter that measures the diverging gap between categories (iii) and (i) with age.
- (b) Modify the model to specify that the slopes of the regression lines with age are equal in categories (i) and (ii) but possibly different in category (iii). Again identify a single parameter measuring divergence.

This and other studies found evidence that the mean depression is high in the late teens, declines toward middle age, and then increases towards old age. Construct a multiple linear regression model in which the association has these characteristics, with possibly different structures in the three education categories. Can this be done in such a way that a single parameter characterizes the divergence hypothesis?

Data Problems

- 20. Kentucky Derby.** Display 9.20 is a partial listing of data in file ex0920 on Kentucky Derby horse race winners from 1896 to 2011. In all those years the race was 1.25 miles in length so that winning time and speed are exactly inversely related. Nevertheless, a simple regression model for changes over time—such as a straight line model that includes *Year* or a quadratic curve that includes *Year* and *Year*²—might work better for one of these response variables than the other. (a) Find a model for describing the mean of either winning time or winning speed as a function of year, whichever works better. (b) Quantify the amount (in seconds or miles per hour) by which the mean winning time or speed on fast tracks exceeds the mean on slow tracks (using the two-category variable *Conditions*), after accounting for the effect of year. (c) After accounting for the effects of year and track conditions, is there any evidence that the mean winning time or speed depends on number of horses in the race (*Starters*)? Is there any evidence of an interactive effect of *Starters* and *Conditions*;

DISPLAY 9.21

Dry weight (mg), ingestion rates (mg per day), and percentage of organic matter in the food, for 22 species of aquatic deposit feeders

Species	Weight	Ingestion	Organic
<i>Hydrobia neglecta</i>	0.20	0.57	18.0
<i>Hydrobia ventrosa</i>	0.20	0.86	17.0
<i>Tubifex tubifex</i>	0.27	0.43	29.7
<i>Hyalella azteca</i>	0.32	0.43	50.0
<i>Potamopyrgus jenkinsi</i>	0.46	2.70	14.4
<i>Hydrobia ulvae</i>	0.90	0.67	13.0
<i>Nereis succinea</i>	5.80	20.20	6.8
<i>Pteronarcys scotti</i>	8.40	1.49	93.0
<i>Orchestia grillus</i>	12.40	4.40	88.0
<i>Arenicola grubii</i>	20.40	240.00	2.2
<i>Thoracophelia mucronata</i>	40.00	230.00	1.0
<i>Ilypolax pusilla</i>	53.00	300.00	4.2
<i>Uca pubnax</i>	63.30	19.90	51.0
<i>Scopimera globosa</i>	65.00	50.00	23.6
<i>Pectinaria gouldii</i>	80.00	1,667.00	0.7
<i>Abarenicola pacifica</i>	380.00	3,400.00	1.2
<i>Abarenicola claparedi</i>	380.00	9,400.00	0.4
<i>Arenicola marina</i>	930.00	4,700.00	0.6
<i>Macrophthalmus japonicus</i>	2,050.00	4,680.00	2.1

that is, does the effect of number of horses on the response depend on whether the track was fast or slow? Describe the effect of number of horses on mean winning time or speed. (Data from Kentucky Derby: Kentucky Derby Racing Results, <http://www.kentuckyderby.info/kentuckyderby-results.php> (July 21, 2011).)

21. Ingestion Rates of Deposit Feeders. Aquatic deposit feeders are organisms that feed on organic material in the bottoms of lakes and oceans. Display 9.21 shows the typical dry weight in mg, the typical ingestion rate (weight of food intake per day for one animal) in mg/day, and the percentage of the food that is composed of organic matter for 19 species of deposit feeders. The organic matter is considered the “quality” part of their food. Zoologist Leon Cammen wondered if, after accounting for the size of the species as measured by the average dry weight, the amount of food intake is associated with the percentage of organic matter in the sediment. If so, that would suggest that either the animals have the ability to adjust their feeding rates according to some perception of food “quality” or that species’ ingestion rates have evolved in their particular environments. Analyze the data to see whether the distribution of species’ ingestion rates depends on the percentage of organic matter in the food, after accounting for the effect of species weight. Also describe the association. Notice that the values of all three variables differ by orders of magnitude among these species, so that log transformations are probably useful. (Data from L. M. Cammen, “Ingestion Rate: An Empirical Model for Aquatic Deposit Feeders and Detritivores,” *Oecologia*, 44 (1980): 303–10.)

22. Mammal Lifespans. The Exercise 8.26 data set contains the mass (in kilograms), average basal metabolic rate (in kilojoules per day), and lifespan (in years) for 95 mammal species. Is metabolic rate a useful predictor of lifespan after accounting for the effect of body mass? What

DISPLAY 9.22

First five rows of a data set with AFQT intelligence test score percentile, years of education achieved by time of interview in 2006, and 2005 annual income in dollars for 1,306 males and 1,278 females in the NLSY survey

Subject	Gender	AFQT	Educ2006	Income2005
2	female	6.841	12	5,500
6	male	99.393	16	65,000
7	male	47.412	12	19,000
8	female	44.022	14	36,000
9	male	59.683	14	65,000

percentage of variation in lifespans (on the log scale) is explained by the regression on log mass alone? What additional percentage of variation is explained by metabolism? Describe the dependence of the distribution of lifespan on metabolic rate for species of similar body mass.

23. Comparing Male and Female Incomes, After Accounting for Education and IQ. Display 9.22 shows the first five rows of a subset of the National Longitudinal Study of Youth data (see Exercise 2.22) with annual incomes in 2005, intelligence test scores (AFQT) measured in 1981, and years of education completed by 2006 for 1,306 males and 1,278 females who were between the ages of 14 and 22 when selected for the survey in 1979, who were available for re-interview in 2006, and who had paying jobs in 2005. Is there any evidence that the mean salary for males exceeds the mean salary for females with the same years of education and AFQT scores? By how many dollars or by what percent is the male mean larger?

Answers to Conceptual Exercises

1. (a) Let $early = 1$ if $time = 24$ and 0 if $time = 0$. Then $\mu\{flowers \mid light, early\} = \beta_0 + \beta_1 light + \beta_2 early$. (b) $\beta_0 + \beta_1 light + \beta_2 early + \beta_3(light \times early)$.
2. The difference is $300 \mu\text{mol}/\text{m}^2/\text{sec}$ times the coefficient of $light$, or about -12.15 flowers.
3. (a) The principal reason is that the 10 plants were all treated together and grown together in the same chamber. The experimental unit is always defined as the unit that receives the treatment, here, plants in the same chamber. (b) The assumption of normality is assisted. Averages tend to have normal distributions, so the averaging may alleviate some distributional problems that could arise from looking at separate numbers of flowers.
4. No. The difficulty with interpreting regression coefficients individually, as in a controlled experiment, is that explanatory variables cannot be manipulated individually. In this instance, the sloth and the fruit bat also have different body weights—the sloth weighs 50 times what the fruit bat weighs. (The full model estimates the brain weight of the fruit bat to be only about 35% of the brain weight of the sloth.) One might attempt to envision a fruit bat having the same weight (0.9 kg) as the sloth and the same litter size (1.0), but having a gestation period of 165 instead of 145 days. This approach, however, is generally unsatisfactory because it extrapolates beyond the experience of the data set (resulting in animals like a fish-eating kangaroo with wings).
5. Yes. A common way to explore lack of fit is to introduce curvature and interaction terms to see if measured effects change as the configuration of explanatory variables changes.
6. Yes.
7. Keep your eye on the parameters. If the mean is linear in the parameters, the model is *linear*.

- (a) Yes, even though it is not linear in X .
(b) Yes.
(c) No. Both numerator and denominator are linear in parameters and X , but the whole is not.
(d) No. This is a very useful model, however.
8. In both, σ is a measure of the magnitude of the difference between a response and the mean in the population from which the response was drawn. In the meadowfoam problem, σ measures the typical size of differences between seedling flowers (averaged from 10 plants) and the mean seedling flowers (averaged from 10 plants) treated similarly (same intensity and timing potential). In the brain weight problem, it is more difficult to describe what σ measures because the theoretical model invents a hypothetical subpopulation of animal species all having the same body weight, gestation, and litter size.
9. (a) $\mu\{pollen \mid duration, queen\} = \beta_0 + \beta_1 duration + \beta_2 queen + \beta_3 (duration \times queen)$, where $queen$ is 1 for queens and 0 for workers (or the other way around). (b) $H_0: \beta_3 = 0$.
10. (a) 8.3 points. (b) As long as the effects are indeed linear in these coded variables, it is a useful way to include them in the multiple regression (and more powerful than considering them to be factors). (c) Yes. The evidence is overwhelming that this coefficient is not zero. (d) The use of premature babies who all had to be fed by tube makes it so that some babies received breast milk but all babies were administered their milk in exactly the same way. (e) It is possible that the decision to provide breast milk and the ability to express breast milk are related to social class and mother's education, which are likely to be related to child's IQ score. It is desired to examine the effect of breast milk that is separate from the association with these potentially confounding variables. (f) (i) 4.5 points. (ii) An important confounding variable in the previous result is the mother's decision to provide breast milk, which may be associated with good mothering skills, which may be associated with better intelligence development in the child. Using the proportion of breast milk as an explanatory variable allows the dose-response assessment of breast milk, which indicates that children of mothers who provided breast milk for 100% of the diet tended to score higher on the IQ test than children of mothers who also decided to provide breast milk but were only capable of supplying a smaller proportion of the diet.
11. (a) 1.93 years. 2.92 years. (b) No. No cause-and-effect is implied by the analysis of these observational data. (c) Seven indicator variables can be included to distinguish the eight graveyards.

Inferential Tools for Multiple Regression

Data analysis involves finding a good-fitting model whose parameters relate to the questions of interest. Once the model has been established, the questions of interest can be investigated through the parameter estimates, with uncertainty expressed through p -values, confidence intervals, or prediction intervals, depending on the nature of the questions asked.

The primary inferential tools associated with regression analysis— t -tests and confidence intervals for single coefficients and linear combinations of coefficients, F -tests for several coefficients, and prediction intervals—are described in this chapter. As usual, the numerical calculations for these tools are performed with the help of a computer. The most difficult parts of the task are knowing what model and what inferential tool best suit the need, and knowing how to interpret and communicate the results.

The inferential tools are illustrated in this chapter on models that incorporate special explanatory variables from the previous chapter: indicator variables, quadratic terms, and interaction terms. The tests and confidence statements found in the examples, and their interpretations, are typical of the ones used in many fields and for many different kinds of data.

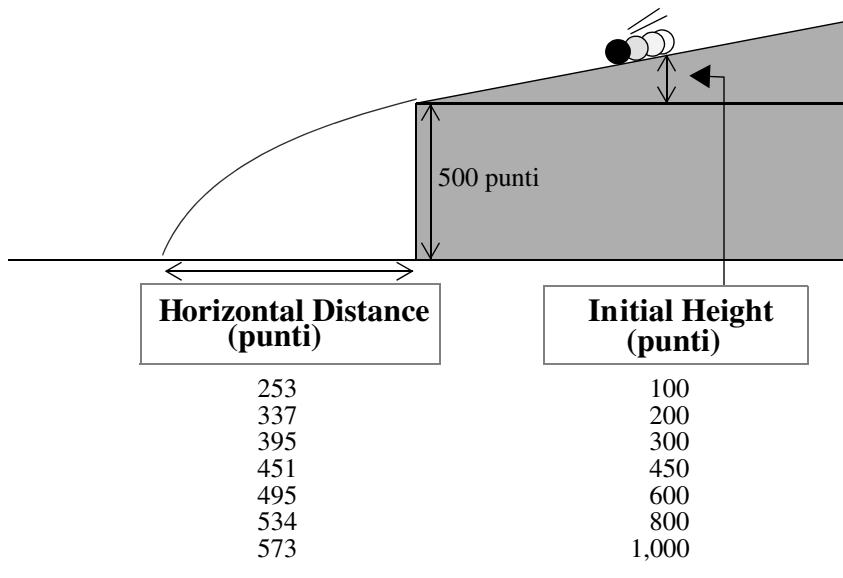
10.1 CASE STUDIES

10.1.1 Galileo's Data on the Motion of Falling Bodies—A Controlled Experiment

In 1609 Galileo proved mathematically that the trajectory of a body falling with a horizontal velocity component is a parabola. His discovery of this result, which preceded the mathematical proof by a year, was the result of empirical findings in an experiment conducted for another purpose.

Galileo's search for an experimental setting in which horizontal motion was not affected appreciably by friction (to study inertia) led him to construct an apparatus like the one shown in Display 10.1. He placed a grooved, inclined plane on a table, released an ink-covered bronze ball in the groove at one of several heights above the table, and measured the horizontal distance between the table and the resulting ink spot on the floor. The data from one experiment are shown in Display 10.1 in units of *punti* (points). One *punto* is 169/180 millimeters. (Data from S. Drake and J. MacLachlan, "Galileo's Discovery of the Parabolic Trajectory," *Scientific American* 232 (1975): 102–10.)

DISPLAY 10.1 Galileo's experimental results



Galileo conducted this experiment to determine whether, in the absence of any appreciable resistance, the horizontal velocity of a moving object is constant. While sketching the paths of the trajectories in his notebook, he apparently came to believe that the trajectory was a parabola. Once the idea of a parabola suggested itself to Galileo, he found that proving it mathematically was straightforward.

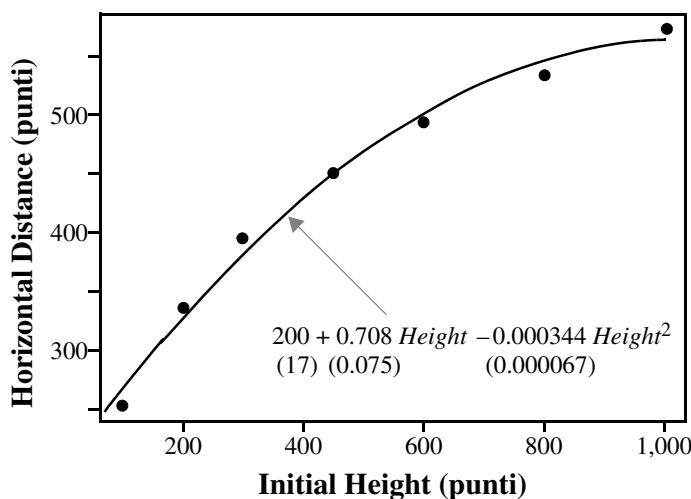
Although Galileo's experiment preceded Gauss's invention of least squares and Galton's empirical fitting of a regression line by more than 200 years, it is interesting to use regression here to explore the regression of horizontal distance on initial height.

Statistical Conclusion

As shown in Display 10.2, a quadratic curve for the regression of horizontal distance on height fits well for initial heights less than 1,000 punti. The data provide strong evidence that the coefficient of a cubic term differs from zero (two-sided p -value = 0.007). Nonetheless, the quadratic model accounts for 99.03% of the variation in measured horizontal distances, and the cubic term explains only an additional 0.91% of the variation. (*Note:* The significance of the cubic term can be explained by the effect of resistance.)

DISPLAY 10.2

Scatterplot of Galileo's horizontal distances versus initial heights, with estimated quadratic regression model (with standard errors in parentheses)



10.1.2 The Energy Costs of Echolocation by Bats—An Observational Study

To orient themselves with respect to their surroundings, some bats use echolocation. They send out pulses and read the echoes that are bounced back from surrounding objects. Such a trait has evolved in very few animal species, perhaps because of the high energy costs involved in producing pulses. Because flight also requires a great deal of energy, zoologists wondered whether the combined energy costs of echolocation and flight in bats was the sum of the flight energy costs and the at-rest echolocation energy costs, or whether the bats had developed a means of echolocation in flight that made the combined energy cost less than the sum.

DISPLAY 10.3

Mass and in-flight energy expenditure for 4 non-echolocating bats (Type = 1), 12 non-echolocating birds (Type = 2), and 4 echolocating bats (Type = 3)

Species	Mass (g)	Type	Flight energy expenditure (W)
<i>Pteropus gouldii</i>	779	1	43.7
<i>Pteropus poliocephalus</i>	628	1	34.8
<i>Hypsignathus monstrosus</i>	258	1	23.3
<i>Eidolon helvum</i>	315	1	22.4
<i>Meliphaga virescens</i>	24.3	2	2.46
<i>Melipsittacus undulatus</i>	35	2	3.93
<i>Sturnus vulgaris</i>	72.8	2	9.15
<i>Falco sparverius</i>	120	2	13.8
<i>Falco tinnunculus</i>	213	2	14.6
<i>Corvus ossifragus</i>	275	2	22.8
<i>Larus atricilla</i>	370	2	26.2
<i>Columba livia</i>	384	2	25.9
<i>Columba livia</i>	442	2	29.5
<i>Columba livia</i>	412	2	43.7
<i>Columba livia</i>	330	2	34.0
<i>Corvus cryptoleucus</i>	480	2	27.8
<i>Phyllostomas hastatus</i>	93	3	8.83
<i>Plecotus auritus</i>	8	3	1.35
<i>Pipistrellus pipistrellus</i>	6.7	3	1.12
<i>Plecotus auritus</i>	7.7	3	1.02

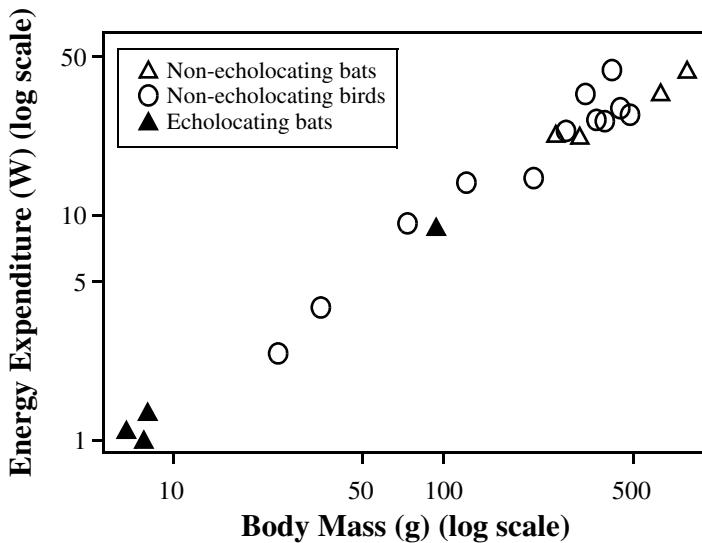
Zoologists considered the data in Display 10.3 on in-flight energy expenditure and body mass from 20 energy studies on three types of flying vertebrates: echolocating bats, non-echolocating bats, and non-echolocating birds. They believed that if the combined energy expenditure for flight and echolocation were additive, the amount of energy expenditure (after accounting for body size) would be greater for echolocating bats than for non-echolocating bats and non-echolocating birds. Display 10.4 shows a log-log scatterplot of in-flight energy expenditure versus body mass for the three types. (Data from J. R. Speakman and P. A. Racey, “No Cost of Echolocation for Bats in Flight,” *Nature* 350 (1991): 421–23.)

Statistical Conclusion

The median in-flight energy expenditure for echolocating bats is estimated to be 1.08 times as large as that for non-echolocating bats, after accounting for body mass. A 95% confidence interval for the ratio of the echolocating median to the non-echolocating median, accounting for body mass, is 0.70 to 1.66. The data are consistent with the hypothesis of equal median in-flight energy expenditures in the three groups, after accounting for body size (p -value = 0.66, extra-sum-of-squares F -test).

DISPLAY 10.4

Log-log scatterplot of in-flight energy expenditure versus body mass for 4 non-echolocating bats, 12 non-echolocating birds, and 4 echolocating bats



Scope of Inference

The species used in the statistical analysis were those for which relevant data were available; any inference to a larger population of species is speculative. The statistical inferences must also be interpreted in light of the likely violation of the independence assumption, due to treating separate studies on the same species as independent observations.

The scientific results are premised on several facts. Bats emit the echolocation pulses once on each wing beat, starting in the latter phase of the up-stroke, coinciding with the moment that air is expelled from the lungs. The coupling of these three processes apparently accounts for the relative energy economy. The energy used to expel air is also used to send out the pulses, and these events occur just before the greatest demand is made on the wing beat.

10.2 INFERENCES ABOUT REGRESSION COEFFICIENTS

Whether the energy expenditure is the same for echolocating bats as for non-echolocating bats, after accounting for body mass, can be investigated by a test of the coefficient of the indicator variable for echolocating bats in the parallel regression lines model. The simplicity with which the question of interest can be investigated makes this the natural choice for an inferential model. Many regression problems share this feature: questions of interest can be answered through tests or confidence intervals for single coefficients in the model.

10.2.1 Least Squares Estimates and Standard Errors

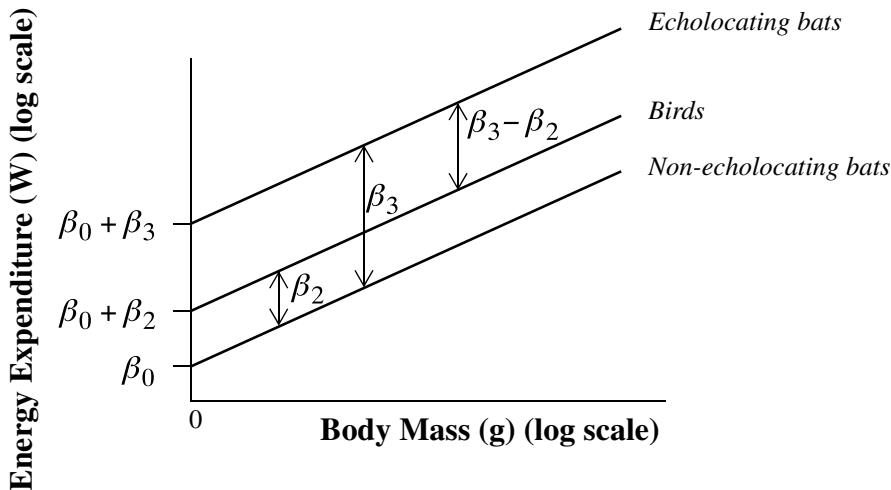
The least squares estimates are the β values that minimize the sum of squared residuals. Formulas for the estimates come from calculus and are best expressed in matrix notation (see Exercises 10.20 and 10.21). A statistical computer program can do these computations. Users can perform all the functions of regression analysis on their data without knowing the relevant formulas. As an example of computer-provided estimates, consider the parallel regression lines model used for the bat echolocation data:

$$\mu\{lenergy \mid lmass, TYPE\} = \beta_0 + \beta_1 lmass + \beta_2 bird + \beta_3 ebat,$$

where $lenergy$ is the log of the in-flight energy, $lmass$ is the log of the body mass, and $TYPE$ is the three-level factor represented by the indicator variables $bird$ (which takes on a value of 1 if the type of species is a bird, and 0 if not) and $ebat$ (which takes on a value of 1 if the species is an echolocating bat, and 0 if not). The third level of $TYPE$, non-echolocating bats, is treated here as the reference level, so its indicator variable does not appear in the regression model. The model is sketched in Display 10.5.

DISPLAY 10.5

The parallel regression lines model for the bat echolocation data



A portion of the output is shown in Display 10.6. The least squares estimates of the β 's appear in the column labeled "Coefficient." The estimate of σ^2 is the sum of squared residuals divided by the degrees of freedom associated with the residuals. The usual rule for finding degrees of freedom applies: (Sample size) minus (Number of unknown regression coefficients in the model for the mean). Since the sample size is 20 and there are 4 coefficients here— β_0 , β_1 , β_2 , and β_3 —the value for degrees of freedom is 16. The square root of the estimate of variance is the

DISPLAY 10.6

Partial summary of the least squares fit to the regression of log energy expenditure on log body mass, an indicator variable for bird, and an indicator variable for echolocating bat

Variable	Coefficient	Standard error	t-statistic	Two-sided p-value
Constant	-1.5764	0.2872	-5.4880	<0.0001
<i>lmass</i>	0.8150	0.0445	18.2966	<0.0001
<i>bird</i>	0.1023	0.1142	0.8956	0.3837
<i>ebat</i>	0.0787	0.2027	0.3881	0.7030

Estimate of $\sigma = 0.1860$, d.f. = 16

estimated standard deviation about the regression, which is 0.1860. This goes by several names on computer output, including *residual SD*, *residual SE*, and *root mean squared error*.

The column labeled “Standard Error” contains the estimated standard deviations of the sampling distributions of the least squares estimates. The formulas for the standard deviations are once again available from statistical theory. In this instance, the formulas can be conveniently expressed only by using matrix notation, as shown in Exercise 10.22. Again, as with simple regression, they depend on the known values of the explanatory variables in the data set and on the unknown value of σ^2 . The standard errors are the values of these standard deviations when σ^2 is replaced by its estimated value. Consequently, the degrees of freedom associated with a standard error are the same as the degrees of freedom associated with the estimate of σ (the sample size minus the number of β 's).

Let $\hat{\beta}_j$ (“beta-hat j ”) represent the estimate of the j th coefficient ($j = 0, 1, 2$, or 3 in the current example). If the distribution of the response for each combination of the explanatory variables is normal with constant variance, then the *t*-ratio,

$$t\text{-ratio} = (\hat{\beta}_j - \beta_j) / \text{SE}(\hat{\beta}_j),$$

has a sampling distribution described by the *t*-distribution with degrees of freedom equal to the degrees of freedom associated with the residuals. This theory leads directly to tests and confidence intervals for individual regression coefficients, in the familiar way.

10.2.2 Tests and Confidence Intervals for Single Coefficients

Do Non-Echolocating Bats Differ from Echolocating Bats?

Test for $\beta_3 = 0$

In the parallel regression lines model for log energy on log mass, the mean log energy for non-echolocating bats is $\beta_0 + \beta_1 lmass$. The mean log energy for echolocating bats is $\beta_0 + \beta_1 lmass + \beta_3$. Thus, for any given mass, the mean log energy expenditure for echolocating bats is β_3 units more than the mean log energy for non-echolocating bats (see Display 10.5). The question of whether the mean log energy expenditure

for echolocating bats is the same as the mean log energy expenditure for non-echolocating bats of similar size may be examined through a test of the hypothesis that β_3 equals 0. From Display 10.6, the two-sided p -value is .7030, providing no reason to doubt that β_3 is zero. The p -value was obtained by the computer as the proportion of values in a t -distribution on 16 degrees of freedom that are farther from 0 than 0.3881 (the t -statistic for the hypothesis that β_3 is 0).

Tests of hypotheses $H: \beta_j = c$ for values of c other than zero are occasionally of interest. For these, the user must construct the t -statistic $(\hat{\beta}_j - c)/SE(\hat{\beta}_j)$ manually and find the p -value by reference to the appropriate t -percentiles.

How Much More Flight Energy Is Expended by Echolocating Bats? Confidence Interval for β_3

Although a test has revealed no reason to doubt that β_3 is zero, neither has it proved that β_3 is zero. Echolocating bats might have a higher energy expenditure, but the available study may not be powerful enough to detect this difference. As is usual for tests of hypotheses, a confidence interval should be reported in addition to the p -value, to emphasize this possibility and to provide an entire set of likely values for β_3 .

The estimate of β_3 is 0.0787, with a standard error of 0.2027. The 97.5th percentile of a t -distribution with 16 degrees of freedom is 2.120, so the 95% confidence interval for β_3 is

$$0.0787 - 2.12(0.2027) \quad \text{to} \quad 0.0787 + 2.12(0.2027),$$

or

$$-0.351 \text{ to } +0.508.$$

Coefficients after log transformations are interpreted in the same way as are coefficients described for simple regression in Section 8.4. If the parallel regression lines model is correct, the median in-flight energy expenditure is $\exp(\beta_3)$ times as great for echolocating bats as it is for non-echolocating bats of similar body mass. A 95% confidence interval for $\exp(\beta_3)$ is obtained by taking the antilogs of the endpoints: $\exp(-0.351) = 0.70$ to $\exp(0.508) = 1.66$. The result can be communicated as follows: It is estimated that the median in-flight energy expenditure for echolocating bats is 1.08 times as great as the median expenditure for non-echolocating bats of similar body size. A 95% confidence interval for this multiplicative effect, accounting for body size, is 0.70 to 1.66.

Significance Depends on What Other Explanatory Variables Are Included

The meaning of the coefficient of an explanatory variable depends on what other explanatory variables are included in the regression. The p -value for the test of whether a coefficient is zero must also be interpreted in light of what other variables are included.

Consider three multiple linear regression models for the echolocation study. The first model says that the mean log energy is different for the three types of flying

animals, but that it does not involve the body weight: $\mu\{lenergy \mid lmass, TYPE\} = TYPE$. The second model—the parallel lines model—indicates that the regression has the same slope against log body weight in all three groups, but different intercepts: $\mu\{lenergy \mid lmass, TYPE\} = lmass + TYPE$. The third model—the separate lines regression model—allows the three groups to have completely different straight line regressions of log energy on log body weight: $\mu\{lenergy \mid lmass, TYPE\} = lmass + TYPE + lmass \times TYPE$. Least squares fits to these three models yield the following results:

- (1) $\hat{\mu}\{lenergy \mid lmass, TYPE\} = 3.40 - 2.74ebat - 0.61bird$
 $(0.42) \quad (0.60) \quad (0.49)$
- (2) $\hat{\mu}\{lenergy \mid lmass, TYPE\} = -1.58 + 0.08ebat + 0.10bird + 0.815lmass$
 $(0.29) \quad (0.20) \quad (0.11) \quad (0.045)$
- (3) $\hat{\mu}\{lenergy \mid lmass, TYPE\} = -0.20 - 1.27ebat - 1.38bird + 0.59lmass$
 $(1.26) \quad (1.29) \quad (1.30) \quad (0.21)$
 $+ 0.21(ebat \times lmass) + 0.25(bird \times lmass)$
 $(0.22) \quad (0.21)$

where the parenthesized numbers beneath the coefficients are their standard errors.

The coefficient of the indicator variable *ebat* is -2.74 in (1), $+0.08$ in (2), and -1.27 in (3). These might appear to be contradictory findings, since the variable is highly significant in (1) but not significant in (2) and (3), and since the sign of the estimated coefficient differs in the three fits. However, a contradiction arises only if one takes the view that the variable *ebat* plays the same role in each equation, which it does not. In (1), the coefficient of *ebat* measures the difference between mean log energy among echolocating bats and mean log energy among non-echolocating bats, ignoring any explanation of differences based on body size (*lmass* is not in the equation to provide the control). It happened (as indicated in Display 10.3) that the echolocating bats were much smaller than the non-echolocating bats in this study. Without taking body size into account, the statistical model attributes all the difference to group differences.

In (2), however, the coefficient of *ebat* measures the difference between energy expenditure of echolocating bats and of non-echolocating bats after adjusting for body size. Interpreting the coefficient as the difference between energy expenditure of echolocating bats and non-echolocating bats “of about the same size” exceeds the scope of this study, because all the echolocating bats were small and all the non-echolocating bats were large. The echolocating versus non-echolocating difference is thus confounded with the differences based on body size. Because “after adjusting for body size” means that group differences are considered only after the best explanation for body size is taken into account, and because body size explains the differences well, it is easy to understand why the coefficient of *ebat* in (2) can be insignificant even though in (1) it is significant.

In (3) the coefficient of *ebat* measures the difference between the intercept parameter in the regression of echolocating bat log energy versus log body weight and the intercept parameter in the regression of non-echolocating bat log energy

versus log body weight. This is a very different interpretation from the one in (2) because here the slopes of the regression equations are allowed to be different. Because the coefficient measures a different quantity in (3) than in (2), there is no reason to expect its statistical significance to be the same.

10.2.3 Tests and Confidence Intervals for Linear Combinations of Coefficients

In the parallel lines model for the echolocation study (represented in Display 10.5), the slope in the regression of log energy on log mass is β_1 for all three groups. The intercept for non-echolocating bats is β_0 , the intercept for non-echolocating birds is $\beta_0 + \beta_2$, and the intercept for echolocating bats is $\beta_0 + \beta_3$.

Since the bird line coincides with the non-echolocating bat line if β_2 is 0, a test of the equality of energy distribution in birds and non-echolocating bats, after accounting for body size, is accomplished through a test of whether β_2 equals 0. Similarly, a test of whether the echolocating bat regression line coincides with the non-echolocating bat regression line is accomplished with a test of $\beta_3 = 0$. The bird and echolocating bat regression lines coincide when $\beta_2 = \beta_3$, however, so a test of the equality of these two groups involves the hypothesis that $\beta_3 - \beta_2 = 0$ (or equivalently, that $\beta_2 - \beta_3 = 0$). This is a hypothesis about a linear combination of regression coefficients.

One method of performing the test compares the estimate of $\beta_3 - \beta_2$ with its standard error. Calculating the standard error is a problem here, however, because the estimates are not statistically independent. The correct formula expands the formula for linear combinations encountered in Section 6.2.2 to include the variances and *covariances* in the joint sampling distribution of the regression coefficient estimates. The correct formula appears in Section 10.4.3. For this problem—as for many similar problems—there is an easier solution.

Redefining the Reference Level

Recall that the choice of reference level for any categorical explanatory variable is arbitrary. Thus, any of the three types of flying vertebrates could be used as the reference level. It is a simple matter to refit the model with another choice for the reference level, and either the birds or the echolocating bats will do. The model will remain the same as the one fit in Display 10.6, although the names of the intercepts will change, and the comparison of the non-echolocating birds to the echolocating bats can be accomplished through a single coefficient—the test for which appears in the standard output.

Inference About the Mean at Some Combination of X's

One special linear combination is the mean of Y at some combination of the X 's. For Galileo's data and the regression model

$$\mu\{distance \mid height\} = \beta_0 + \beta_1 height + \beta_2 height^2,$$

the mean distance at a height of 250 is

$$\mu\{distance \mid height = 250\} = \beta_0 + (\beta_1 \times 250) + (\beta_2 \times (250)^2),$$

which is estimated by the linear combination of the estimates

$$\hat{\mu}\{distance \mid height = 250\} = (\hat{\beta}_0 \times 1) + (\hat{\beta}_1 \times 250) + (\hat{\beta}_2 \times (250)^2).$$

A standard error for this estimate may be obtained by using the methods of Section 10.4.3, but it may also be extracted from a computer analysis, by redefining the reference level for height. The trick is to create a new explanatory variable by subtracting 250 from each height measurement. Let $ht250 = height - 250$, so $height = 250$ corresponds to $ht250 = 0$, and then fit the model

$$\mu\{distance \mid ht250\} = \beta_0^* + \beta_1^*ht250 + \beta_2^*(ht250)^2.$$

The mean distance at $height = 250$ is $\mu\{distance \mid ht250 = 0\} = \beta_0^*$. Redefining the model terms permits the question of interest to be worded in terms of a single parameter; the standard error for this parameter appears in the usual output. Use the estimated intercept as the estimated mean of interest, and use the reported standard error of the intercept as the standard error for this estimated mean. As with the indicator variable example, the two models are identical, but are parameterized differently. Display 10.7 shows the analysis for both ways of writing the quadratic regression model, emphasizing how to obtain both the estimate of the mean (at $height = 250$ in this example) and its standard error.

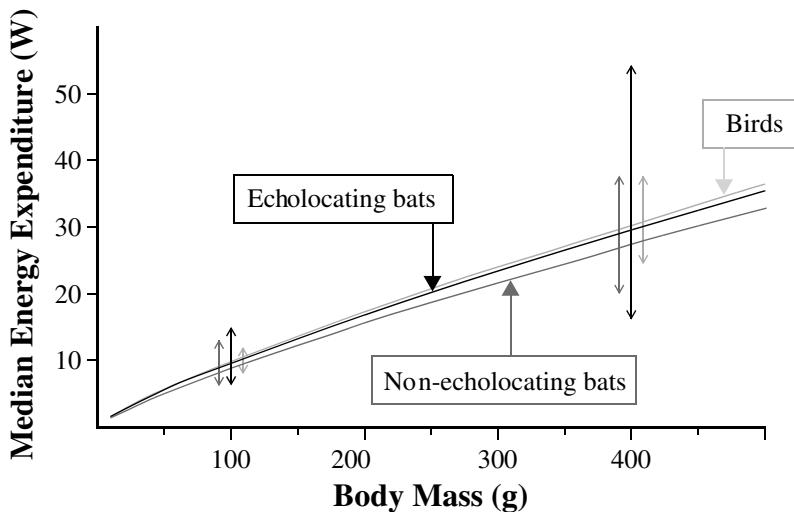
DISPLAY 10.7

Estimates of polynomial coefficients with two different reference levels of height, in Galileo's study

(Reference $height = 0$ punti)						
Variable	Coefficient	Standard error	t-statistic	Two-sided p-value		
Constant	199.91	16.76	11.93	0.0003		
$height$	0.7083	0.0748	9.47	0.0007		
$height^2$	-0.0003437	0.0000668	5.15	0.0068		
R -squared = 99.0%		Adj. R -squared = 98.6%	Estimated SD = 13.6 punti			
$\hat{\mu}\{distance \mid height = 250 \text{ punti}\}$ $SE(\hat{\mu}\{distance \mid height = 250 \text{ punti}\})$						
Reference $height = 250$ punti						
Variable	Coefficient	Standard error	t-statistic	Two-sided p-value		
Constant	355.51	6.625	53.66	<0.0001		
$height - 250$	0.5365	0.0430	12.48	0.0002		
$(height - 250)^2$	-0.0003437	0.0000668	-5.15	0.0068		
R -squared = 99.0%		Adj. R -squared = 98.6%	Estimated SD = 13.6 punti			

DISPLAY 10.8

Estimated median energy expenditures for birds, echolocating bats, and non-echolocating bats as functions of body mass (parallel lines model on log-log scale, with 95% confidence bands)



Confidence Bands for Multiple Regression Surfaces

Display 10.8 shows the estimated regressions for the bat data (the parallel regression lines for the log of energy expenditure on the log of body mass) on the original scales of measurement. For non-echolocating bats, for example,

$$\text{Median}\{\text{energy} \mid \text{mass}\} = \exp(\beta_0) \times \text{mass}^{\beta_1}.$$

The curve for this group on the plot was drawn by computing the estimated median at many values of *mass* and connecting the resulting points.

The double arrows in Display 10.8 show the limits of a 95% confidence band for the median energy expenditure at 100 and 400 grams for each of the three groups. These intervals were constructed by finding confidence bands for the regression surface of $\log(\text{energy})$ and then back-transforming the endpoints. The intervals were constructed by using the computer trick for finding the standard errors of estimated means. This required six different fits to the model, with different reference values, as shown in Display 10.9. The estimated mean for birds of 100-gram body mass, for example, was the estimated intercept in the regression of *lenergy* on (*lmass* – 100) and indicator variables for the two groups other than birds. This turned out to be 2.2789, with a standard error of 0.0604. From these, the confidence interval is calculated and then back-transformed for the light gray double-arrow line at Body mass = 100 in Display 10.8. The calculations appear in the bottom portion of Display 10.9.

Notice that the multiplier used in the confidence band is based on the 95th percentile from an *F*-distribution. (This example assumes that a confidence band over the entire region—all masses and types—is desired, rather than confidence

DISPLAY 10.9

Construction of the 95% confidence band, using repeated fits of the multiple regression model with different reference points

(1) *Computer Work*

Reference point		Explanatory variables		
TYPE	Body mass	TYPE indicators	Body mass variable	Intercept estimate
birds	100	<i>nbat, ebat</i>	<i>lmass-log(100)</i>	2.2789
	400	"	<i>lmass-log(400)</i>	3.4087
non-echo bats	100	<i>ebat, bird</i>	<i>lmass-log(100)</i>	2.1767
	400	"	<i>lmass-log(400)</i>	3.3064
echo bats	100	<i>nbat, bird</i>	<i>lmass-log(100)</i>	2.2553
	400	"	<i>lmass-log(400)</i>	3.3851

(2) *Hand Calculations—an Example*

$$\text{Multiplier} = \sqrt{4 F_{4,16; 0.95}} = 3.468$$

$$\text{Lower limit} = \exp[2.2789 - (3.468)(0.0604)] = 7.9$$

$$\text{Upper limit} = \exp[2.2789 + (3.468)(0.0604)] = 12.0$$

intervals at a few specific values.) This type of multiplier was first introduced as the Scheffé multiple comparison procedure in Section 6.4.2 and again for confidence bands for simple regression in Section 7.4.2. Here, the numerator and denominator degrees of freedom in F are equal to the number of parameters in model (4) and to the residual degrees of freedom (16), respectively. The multiplier of F inside the square root sign is equal to the number of parameters in the model (4, in this case). Use of the F -based multiplier guarantees at least 95% confidence that the bands contain the regression surface throughout the experimental range. (*Note:* It is dangerous to draw conclusions about the regression outside of the scope of the available data. For example, no non-echolocating bats had body sizes of less than 258 grams. Inference about the distribution of energy expenditure for non-echolocating bats of 100-gram body size requires extrapolation of the model beyond the range within which it was estimated.)

10.2.4 Prediction

Prediction is an important objective in many regression applications. If it is the only objective, there is no need to interpret the coefficients. For example, a researcher may be able to predict individuals' blood pressures from the number of bathrooms in their houses, without having to interpret or understand the relationship between those two variables.

Predicted values are the same as estimated means. They are calculated by evaluating the estimated regression at the desired explanatory variable values, as

in the previous section. As in simple linear regression, the error in prediction comes from two sources: the error of estimating the population mean, and the error inherent in the fact that the new observation differs from its mean. Since $\hat{Y} = \text{pred}\{Y|X\} = \hat{\mu}\{Y|X\}$, the prediction error is

$$Y - \hat{Y} = Y - \hat{\mu}\{Y|X\} = [Y - \mu\{Y|X\}] - [\hat{\mu}\{Y|X\} - \mu\{Y|X\}].$$

The variance of a prediction error consists of two corresponding pieces:

$$\text{Prediction variance} = \sigma^2 + [\text{SD}(\hat{\mu}\{Y|X\})]^2.$$

The variance of prediction at some explanatory variable value, therefore, may be estimated by adding the estimate of σ^2 to the square of the standard error of the estimated mean. (See also Section 7.4.3.) The standard error of prediction is the square root of the estimated variance of prediction. Prediction intervals are based on this standard error and on the appropriate percentile from the t -distribution with degrees of freedom equal to those associated with the estimate of σ . As discussed with regard to simple regression, predictions are valid only within the range of values of explanatory variables used in the study. The validity also depends fairly strongly on the assumption of normal distributions in the populations.

Galileo Example

Display 10.7 has all the information needed to predict the horizontal distance of a single ball released at a height of 250 punti. The predicted distance is 355.5 punti. The variance of prediction is $(13.6)^2 + (6.62)^2$, whose square root is the standard error of prediction = 15.1 punti. With $7 - 3 = 4$ degrees of freedom, the t -multiplier for a 95% prediction interval is 2.776, so the interval extends from $355.5 - 42.0 = 313.5$ punti to $355.5 + 42.0 = 397.5$ punti.

10.3 EXTRA-SUMS-OF-SQUARES F-TESTS

In multiple regression, data analysts often need to test whether *several* coefficients are all zero. For example, the model

$$\mu\{lenergy | lmass, TYPE\} = \beta_0 + \beta_1 lmass + \beta_2 bird + \beta_3 ebat$$

presents the regression as three parallel straight lines. The lines are identical under the hypothesis

$$H: \beta_2 = 0 \text{ and } \beta_3 = 0.$$

The alternative hypothesis is that at least one of the coefficients— β_2 or β_3 —is nonzero. t -tests, either individually or in combination, cannot be used to test such a hypothesis involving more than one parameter.

10.3.1 Comparing Sizes of Residuals in Hierarchical Models

The extra-sum-of-squares method, on the other hand, is ideally suited for this purpose. It directly compares a full model (*lmass + TYPE*) to a reduced model (*lmass*):

$$\begin{aligned}\textbf{Full: } \mu\{lenergy \mid lmass, TYPE\} &= \beta_0 + \beta_1 lmass + \beta_2 bird + \beta_3 ebat, \\ \textbf{Reduced: } \mu\{lenergy \mid lmass, TYPE\} &= \beta_0 + \beta_1 lmass.\end{aligned}$$

If the coefficients in the last two terms of the full model are zero, their estimates should be close to zero, and fits to the two models should give about the same results. In particular, the residuals should be about the same size. If either β_2 or β_3 is not zero, however, the full model should do a better job of explaining the variation in the response variables, and the residuals should tend to be smaller in magnitude than those from the reduced model. Even if the reduced model is correct, however, the squared residuals in the full model must be somewhat smaller, since the full model has more flexibility to match chance variations in the data. The *F*-test is used to assess whether the difference between the sums of squared residuals from the full and reduced models is greater than can be explained by chance variation. The sum of squared residuals in the reduced model minus the sum of squared residuals in the full model (as in Section 5.3) is the *extra sum of squares*:

$$\begin{aligned} \textit{Extra sum of squares} = & \\ \textit{Sum of squared residuals from reduced model} - & \\ \textit{Sum of squared residuals from full model}. &\end{aligned}$$

The sum of squared residuals is a measure of the response variation that remains unexplained by a model. The extra sum of squares may be interpreted as being the amount by which the unexplained variation decreases when the extra terms are added to the reduced model. Or it may be interpreted as being the amount by which the unexplained variation increases when the extra terms are dropped from the full model.

10.3.2 *F*-Test for the Joint Significance of Several Terms

The extra-sum-of-squares *F*-test has been encountered previously as a one-way analysis of variance tool for comparing the separate-means model to the single-mean model (Section 5.3), and as a tool in simple regression for comparing the regression model to the single-mean model or to the separate-means model if there are replicates (Section 8.5). These are all special cases of a general form that is often called a *partial F-test*.

The *F*-statistic, based on the extra sum of squared residuals in a reduced model over a full model is defined as follows:

$$F\text{-statistic} = \frac{\left[\begin{array}{c} \text{Extra sum of squares} \\ \hline \text{Number of betas being tested} \end{array} \right]}{\text{Estimate of } \sigma^2 \text{ from full model}}.$$

The part in brackets is the average size of the extra-sum-of-squares per coefficient being tested. If the reduced model is correct, then this per-coefficient variation should be roughly equal to the per observation variation, σ^2 . The F -statistic, therefore, should be close to 1.

When the reduced model is correct and the rest of the model assumptions (including normality) hold, the sampling distribution of this F -statistic is an F -distribution. (Although exact justification is based on normality, the F -test and the t -tests are robust against departures from normality.) The numerator degrees of freedom are the number of β 's being tested. This is either the number of β 's in the full model minus the number of β 's in the reduced model or, equivalently, the residual degrees of freedom in the reduced model minus the residual degrees of freedom in the full model. The denominator degrees of freedom are those associated with the estimate of σ^2 in the full model. If the F -statistic is larger than expected from this F -distribution, this is interpreted as evidence that the reduced model is incorrect. The test's p -value—the chance that an F -variable exceeds the calculated value of the F -statistic—measures the strength of that evidence.

Example—Bat Echolocation Data

Computations in the extra-sum-of-squares F -test are detailed in Display 10.10, leading to a conclusion that there is no evidence of a group difference.

Special Case: Testing the Significance of a Single Coefficient

One special hypothesis is that a single coefficient is zero; for example, $H: \beta_3 = 0$. Since the t -test is available for this hypothesis, one might wonder whether the F -test leads to the same result. It does. The F -statistic is the square of the t -statistic, and the p -value from the F -test is the same as the two-sided p -value from the t -test. As a practical matter, the t -test results are more convenient to obtain from computer output.

Special Case: Testing the "Overall Significance" of the Regression

Another special hypothesis is that all regression coefficients except β_0 are zero. This hypothesis proposes that none of the considered explanatory variables are useful in explaining the mean response. Although this hypothesis only occasionally corresponds to a question of interest, the extra-sums-of-squares F -test for this hypothesis—often called the F -test for overall significance of the regression—is routine output in most statistical computer programs. As discussed in the following section,

DISPLAY 10.10

The extra-sum-of-squares F -test for testing equality of intercepts in the parallel regression lines model (bat echolocation data)

1

Fit the FULL model: $\mu\{lenergy| lmass, TYPE\} = \beta_0 + \beta_1 lmass + \beta_2 bird + \beta_3 ebat$.

$$\text{Sum of squared residuals} = 0.55332 \quad \text{d.f.} = 16 \quad \hat{\sigma}^2 = 0.03458$$

2

Fit the REDUCED model: $\mu\{lenergy| lmass, TYPE\} = \beta_0 + \beta_1 lmass$.

$$\text{Sum of squared residuals} = 0.58289 \quad \text{d.f.} = 18$$

3

The extra sum of squares is the difference between the two residual sums of squares.

$$\text{Extra SS} = 0.58289 - 0.55332 = 0.02957$$

4

Numerator degrees of freedom are the number of β 's in the full model minus the number of β 's in the reduced model.

$$\text{Numerator d.f.} = 4 - 2 = 2$$

5

Calculate the F -statistic.

$$F\text{-statistic} = \left[\frac{0.02957}{2} \right] = \frac{0.014785}{0.03458} = 0.428$$

6

Find $Pr(F_{2,16} > 0.428)$ from table, computer, or calculator.

$$p\text{-value} = 0.66$$

Conclusion: There is no evidence that mean log energy differs for birds, echolocating bats, and non-echolocating bats, after accounting for body mass.

the component calculations of this F -test are routinely available in an analysis of variance table.

10.3.3 The Analysis of Variance Table

The reduced model for the mean contains only the constant term β_0 . Hence the least squares estimate of β_0 is the average response (\bar{Y}), and the sum of squared residuals is $(n - 1)s_Y^2$. This is the “Total sum of squares” as it measures the total variation of the responses about their average. The sum of squared residuals from the model under investigation is labeled the “Residual sum of squares” (or sometimes the “Error sum of squares”) in the analysis of variance table. The difference between these two—the extra-sum-of-squares—is the amount of total variation in the response variable that can be explained by the regression on the explanatory variables. It is called the “Regression sum of squares” or the “Model sum of squares.” An analysis of variance table for the quadratic fit to Galileo’s data appears in Display 10.11.

DISPLAY 10.11

Analysis of variance table for Galileo's data: fit of the data to the quadratic model: $\mu\{distance | height\} = \beta_0 + \beta_1 height + \beta_2 height^2$ (based on seven observations)

Source of Variation	Sum of Squares	d.f.	Mean Square	F-Statistic	p-Value
Regression	76277.92	2	38138.96	205.03	<0.0001
Residual	744.08	4	186.02		
Total	77022.00	6			

1 Residual sum of squares from reduced model: $Mean\{distance\} = \beta_0$.

2 Residual sum of squares from full model: $Mean\{distance\} = \beta_0 + \beta_1 height + \beta_2(height)^2$.

3 The extra sum of squares is the total sum of squares minus the residual sum of squares (the amount of variation explained by the two explanatory variables).

4 Mean squares are always sums of squares divided by their degrees of freedom. The residual mean square is the estimate of σ^2 .

5 The F-statistic (for overall significance of regression) is the regression mean square divided by the residual mean square.

6 $p\text{-value} = Pr(F_{2,4} > 205.03)$. The small p-value here indicates overwhelming evidence that the coefficient of at least one of the explanatory variables height and $height^2$ differs from zero.

Using Analysis of Variance Tables for Extra-Sums-of-Squares Tests

The calculations for extra-sums-of-squares tests can be carried out manually, as shown in Display 10.10. The sums of squared residuals in steps 1 and 2 are obtained by fitting the full and reduced models, obtaining the analysis of variance tables for each, and reading the sum of squared residuals from the tables. Some statistical computing packages will perform the calculations with an analysis of variance command that includes the full and reduced models as user-specified arguments.

Analysis of the Bat Echolocation Data

The analysis of the bat echolocation data involves several extra-sums-of-squares F-tests. The main question of interest is, “Is there a difference in the in-flight energy expenditures of echolocating and non-echolocating bats after body size is accounted for?” This question does not involve non-echolocating birds, but including them in the analysis is important for obtaining the best possible estimate of σ^2 . A secondary question compares birds to the two bat groups.

Inspection of the data and initial graphical exploration indicated the need to analyze both energy and body mass on the log scale. The statistical analysis starts coming into focus with a graphical display like the coded scatterplot in Display 10.4.

It is apparent that a straight line regression of log energy on log body mass is probably appropriate for each of the three types, and there is no obvious evidence of a difference in the slopes of these lines for the three types of flying vertebrates. These results are fortunate, because the question of interest is most easily addressed if the group differences are the same for all body weights. Hence, the parallel lines model offers the most convenient inferential model.

To buttress the argument, the analyst must make sure that the regression lines are, indeed, parallel before testing whether the intercepts are equal. This involves first fitting a rich model that incorporates different intercepts and different slopes. The separate lines model is

$$\begin{aligned}\mu\{lenergy \mid lmass, TYPE\} = \\ \beta_0 + \beta_1 lmass + \beta_2 bird + \beta_3 ebat + \beta_4(lmass \times bird) + \beta_5(lmass \times ebat).\end{aligned}$$

The next step is to fit the rich model and examine a residual plot (plot of residuals versus fitted values). This plot (not shown) indicates no problems with the model assumptions, so it is appropriate to perform the *F*-test for the hypothesis that the interaction terms can be dropped (namely, $H: \beta_4 = \beta_5 = 0$). This *F*-test, based on the analysis of variance tables, is shown in Display 10.12.

Since the *p*-value is 0.53, it is safe to drop the interaction terms and address the comparison of interest in terms of the parallel lines model. It is then appropriate to test whether the intercepts in the parallel regression lines model are the same for the three groups. This is accomplished by conducting another extra-sum-of-squares *F*-test, as previously illustrated in Display 10.10. The *p*-value of 0.66 implies that the data are consistent with the hypothesis that no difference in energy expenditure exists among the three groups, after the effect of body mass is accounted for. Nevertheless, confidence intervals for the group differences should still be presented, as in the summary of statistical findings in Section 10.1.2, to reveal other possibilities that are not excluded by the data.

10.4 RELATED ISSUES

10.4.1 Further Notes on the *R*-Squared Statistic

Example—Galileo’s Data

From the analysis of variance table in Display 10.11, it is evident that the quadratic regression of *distance* on *height* explains 76277.92/77022.00, or 99.03% of the total variation in the observed distances. Only 0.97% of the variation in distances remains unexplained.

Display 10.13 shows the regression output for another model for these data: the regression of distance on *height*, *height-squared*, and *height-cubed*. The *p*-value for the coefficient of *height-cubed* (0.0072) provides strong evidence that the cubic term is significant even when the linear and quadratic terms are in the model. On the other hand, the percentage of variation explained by the cubic model is 99.94%, only 0.91% more than the percentage of variation explained by the quadratic model.

DISPLAY 10.12

The extra sum of squares F -test comparing the separate regression lines model to the parallel regression lines model—bat echolocation data

- 1** FIT FULL MODEL: $\mu\{lenergy|lmass, TYPE\} = \beta_0 + \beta_1 lmass + \beta_2 bird + \beta_3 ebat + \beta_4(lmass \times bird) + \beta_5(lmass \times ebat)$.

Source of Variation	Sum of Squares	d.f.	Mean Square	F-Statistic	p-Value
Regression	29.46993	5	5.89399	163.4	<0.0001
Residual	0.50487	14	0.03606		
Total	29.97480	19			

Residual SS

Estimate of σ^2
d.f.

- 2** FIT REDUCED MODEL: $\mu\{lenergy|lmass, TYPE\} = \beta_0 + \beta_1 lmass + \beta_2 bird + \beta_3 ebat$.

Source of Variation	Sum of Squares	d.f.	Mean Square	F-Statistic	p-Value
Regression	29.42148	3	9.80716	283.6	<0.0001
Residual	0.55332	16	0.03458		
Total	29.97480	19			

Residual SS

- 3** The extra sum of squares is the difference between residual sums of squares.

$$\text{Extra SS} = 0.55332 - 0.50487 = 0.04845$$

5 Calculate the F-statistic. $\rightarrow F\text{-statistic} = \frac{\left[\frac{0.04845}{2}\right]}{0.03606} = 0.672$

Numerator d.f. = Number of β 's in full model minus Number of β 's in reduced model,

6 Look up $Pr(F_{2,14} > 0.672)$. $\rightarrow p\text{-value} = 0.53$

Conclusion: There is no evidence that the association between energy expenditure and body size differs among the three types of flying vertebrates ($p\text{-value} = 0.53$).

Of course that 0.91% accounts for 94% of the remaining variability from the quadratic model, which helps explain why the cubic term is statistically significant.

Two very useful features of R^2 have been revealed with Galileo's data. First, the summary measure—for instance, 99.03% of the variation explained—provides a valuable image of the tightness of the fit. Second, the proportion of additional

DISPLAY 10.13

Partial output from the regression of distance on *height*, *height-squared*, and *height-cubed*, for Galileo's data

Variable	Coefficient	Standard error	t-statistic	p-value
Constant	155.78	8.33	18.71	0.0003
<i>height</i>	1.1153	0.0657	16.98	0.0004
<i>height-squared</i>	-0.001245	0.000138	-8.99	0.0029
<i>height-cubed</i>	5.477×10^{-7}	0.838×10^{-7}	6.58	0.0072

Estimate of standard deviation about the regression: 4.011 on 3 degrees of freedom
 $R^2 = 99.94\%$.

variation explained by one newly introduced variable after the others have been accounted for can be used in assessing its practical significance. Despite these useful summarizing applications, however, R^2 suffers considerable abuse. It is rarely an appropriate statistic to use for model checking, model comparison, or inference.

***R*² Can Always Be Made 100% by Adding Explanatory Variables**

Display 10.14 shows the scatterplot of distance versus height for Galileo's seven measurements. Drawn through the points is a sixth-order polynomial regression line:

$$\begin{aligned}\mu\{\text{distance} \mid \text{height}\} &= \beta_0 + \beta_1 \text{height} + \beta_2(\text{height})^2 + \beta_3(\text{height})^3 \\ &\quad + \beta_4(\text{height})^4 + \beta_5(\text{height})^5 + \beta_6(\text{height})^6.\end{aligned}$$

This fit produces residuals that are all zero and, consequently, an R^2 of 100%. In fact, for n data points with distinct explanatory variable values, an $n - 1$ polynomial regression will always fit exactly. That does not imply that the equation has any usefulness, though. While the data at hand have been modeled exactly, the particular equation is unlikely to fit as well on future data.

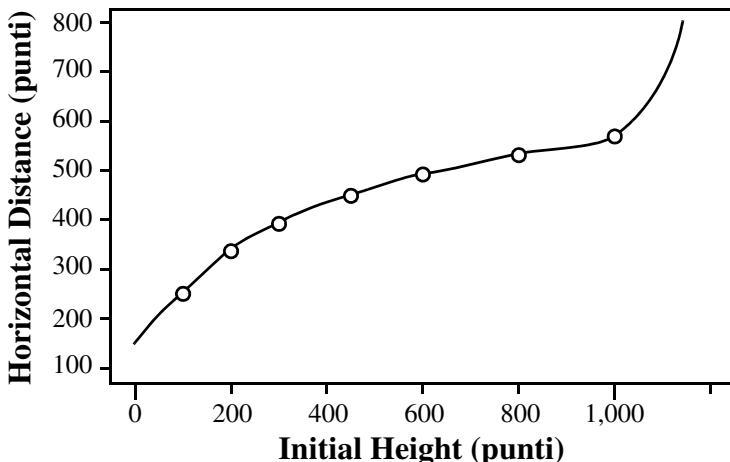
This same warning applies to multiple regression in general. R -squared can always be made 100% by adding enough explanatory variables. The question is whether the model is appropriate for the population or instead is only appropriate for the particular data on hand. If inference to some population (real or hypothetical) is desired, it is important not to craft a model that is overly customized to the data on hand. Tests of significance and residual plots are more appropriate tools for model building and checking.

***The Adjusted R*² Statistic**

The adjusted R^2 is a version of R^2 that includes a penalty for unnecessary explanatory variables. It measures the proportion of the observed spread in the responses that is explained by the regression model, where spread is measured by residual mean squares rather than by residual sums of squares. This approach provides a

DISPLAY 10.14

Scatterplot of Galileo's horizontal distances versus initial heights, with estimated sixth-order polynomial regression curve ($R^2 = 100\%$)



better basis for judging the improvement in a fit due to adding an explanatory variable, but it does not have the simple summarizing interpretation that R^2 has:

$$\text{Adjusted } R^2 = 100 \frac{(\text{Total mean square}) - (\text{Residual mean square})}{\text{Total mean square}} \text{%.}$$

Thus adjusted R^2 is useful for casual assessment of improvement of fit, but R^2 is better for description, as illustrated in the summary of statistical findings in Section 10.1.1.

10.4.2 Improving Galileo's Design with Replication

Since Galileo's study contained no replication, it is impossible to separate the repeatable part of the relationship (if any) from the real variability in replicate measurements of horizontal distances of balls started at the same height. In other words, all estimates of σ are necessarily based on a particular form of the regression model and may be biased as a result of model inadequacies.

Suppose that Galileo instead had selected four evenly spaced heights—say 100, 400, 700, and 1,000 punti—and had replicated the experiment at each height. Using four heights would have allowed Galileo to fit a cubic polynomial, giving him a model on which to judge the proposed parabola (through a test of the cubic term). The revised design also yields 4 degrees of freedom for estimating residual variability free from possible model complications. The cubic model can be judged by comparing it to the separate-means model for the four groups.

A natural, *design-based* estimate of variance is obtained by pooling sample variances from groups of data at identical values of the explanatory variables. It can only be obtained if there are replicates. Without replication, the only available estimate of variance is the one obtained from the residuals about the fit to the presumed model for the mean. There is always a possibility of bias in this *model-based* estimate of variance, due to inadequacies of the regression model. With a design-based estimate no such bias can occur, because the size of the residuals does not hinge on the accuracy of any presumed regression model. This is a reason for incorporating some replication into an experimental design.

10.4.3 Variance Formulas for Linear Combinations of Regression Coefficients

On some occasions it may seem more convenient to estimate a linear combination of regression parameters directly than to instruct the computer to change the reference level (as in Section 10.2.3). In general, a linear combination can be written as

$$\gamma = C_0\beta_0 + C_1\beta_1 + C_2\beta_2 + \cdots + C_p\beta_p,$$

where the C 's are known coefficients, some of which may be zeros. The estimate of this combination is

$$g = C_0\hat{\beta}_0 + C_1\hat{\beta}_1 + C_2\hat{\beta}_2 + \cdots + C_p\hat{\beta}_p.$$

The formula for the variance of this linear combination differs from the one for variance of a linear combination of independent averages in Section 6.2.2. Since the estimated coefficients are not independent of one another, the formula here involves their covariances.

For a population of pairs U and V , the *covariance* of U and V is the mean of $(U - \mu_U)(V - \mu_V)$, and it describes how the two variables covary. The covariance of $\hat{\beta}_1$ and $\hat{\beta}_2$ is the mean of $(\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2)$ in their sampling distribution. A formula for the covariances of least squares estimates is available, and estimates are available from least squares computer routines. If Cov represents covariance, the variance of the sampling distribution of the estimated linear combination is

$$\begin{aligned} \text{Var}\{g\} &= C_0^2\text{SE}(\hat{\beta}_0)^2 + C_1^2\text{SE}(\hat{\beta}_1)^2 + \cdots + C_p^2\text{SE}(\hat{\beta}_p)^2 + 2C_0C_1\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\ &\quad + 2C_0C_2\text{Cov}(\hat{\beta}_0, \hat{\beta}_2) + \cdots + 2C_{p-1}C_p\text{Cov}(\hat{\beta}_{p-1}, \hat{\beta}_p). \end{aligned}$$

To see why this is so consider the linear combination $\gamma = C_1\beta_1 + C_2\beta_2$ and its estimate $g = C_1\hat{\beta}_1 + C_2\hat{\beta}_2$. Then

$$(g - \gamma) = C_1(\hat{\beta}_1 - \beta_1) + C_2(\hat{\beta}_2 - \beta_2)$$

and

$$(g - \gamma)^2 = C_1^2(\hat{\beta}_1 - \beta_1)^2 + C_2^2(\hat{\beta}_2 - \beta_2)^2 + 2C_1C_2(\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2).$$

DISPLAY 10.15

Inference about $\beta_2 - \beta_3$, the coefficient of the indicator variable for birds minus the coefficient of the indicator variable for echolocating bats

1

Estimate the linear combination of coefficients as the same linear combination of estimated coefficients.

Estimate of $\beta_2 - \beta_3$, from Display 10.6: $0.1023 - 0.0787 = 0.0236$

2

Obtain the estimated variance–covariance matrix of the estimated regression coefficients.

Estimated variance–covariance matrix (from computer):

	(Constant)	<i>lmass</i>	<i>bird</i>	<i>ebat</i>
Constant	0.08250	-0.01211	-0.01921	-0.05056
<i>lmass</i>	-0.01211	0.00198	0.00173	0.00687
<i>bird</i>	-0.01921	0.00173	0.01304	0.01464
<i>ebat</i>	-0.05056	0.00687	0.01464	0.04108

Notes: The matrix is symmetric. The square roots of the diagonal elements are the standard errors of the estimated coefficients (as reported in Display 10.6).

3 The estimated variance of $\hat{\beta}_2 - \hat{\beta}_3$ is

$$1^2 \text{Var}(\hat{\beta}_2) + (-1)^2 \text{Var}(\hat{\beta}_3) + 2(1)(-1) \text{Cov}(\hat{\beta}_2, \hat{\beta}_3).$$

$$\text{Estimated variance of } \hat{\beta}_2 - \hat{\beta}_3: 0.01304 + 0.04108 - (2 \times 0.01464) = 0.02484$$

$$\text{SE}(\hat{\beta}_2 - \hat{\beta}_3) = (0.02484)^{1/2} = 0.1576 \quad (\text{d.f.}=16)$$

The variance of g is $\text{Mean}(g - \gamma)^2$, which can be expressed as

$$\begin{aligned} C_1^2 \text{Mean}(\hat{\beta}_1 - \beta_1)^2 + C_2^2 \text{Mean}(\hat{\beta}_2 - \beta_2)^2 + 2C_1C_2 \text{Mean}(\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2) \\ = C_1^2 \text{Var}(\hat{\beta}_1) + C_2^2 \text{Var}(\hat{\beta}_2) + 2C_1C_2 \text{Cov}(\hat{\beta}_1, \hat{\beta}_2). \end{aligned}$$

The estimate of this expression is obtained by substituting the estimates of the variances and covariance.

The Estimated Variance–Covariance Matrix for the Coefficients

The estimated covariances of the coefficients' sampling distributions are calculated by most statistical computer programs. They are stored as an array, or *matrix*, for use in formulas such as the one just discussed. (See also Exercise 10.22.) The easy computer trick for finding the standard error of $\hat{\beta}_2 - \hat{\beta}_3$ for the parallel regression lines model fit to the echolocation data was shown in Section 10.2.3. Display 10.15 illustrates the alternative, direct approach, using the estimated coefficient covariances supplied by a statistical computer program.

10.4.4 Further Notes About Polynomial Regression

A Second-Order Model in Two Explanatory Variables

A full, *second-order model* in explanatory variables X_1 and X_2 is

$$\mu\{Y|X_1, X_2\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2.$$

This is a useful model in *response surface studies*, where the goal is to model a nonlinear response surface in terms of a polynomial. It is rarely useful to attempt cubic and higher-order terms.

When Should Squared Terms Be Included?

As with interaction terms, quadratic terms should not routinely be included. They are useful to consider in four situations: when the analyst has good reason to suspect that the response is nonlinear in some explanatory variable (through knowledge of the process or by graphical examination); when the question of interest calls for finding the values that maximize or minimize the mean response; when careful modeling of the regression is called for by the questions of interest (and presumably this is only the case if there are just a few explanatory variables); or when inclusion is used to produce a rich model for assessing the fit of an inferential model.

10.4.5 Finding Where the Mean Response Is at Its Maximum in Quadratic Regression

The value of X that maximizes (or minimizes) the quadratic expression

$$\mu\{Y|X\} = \beta_0 + \beta_1 X + \beta_2 X^2$$

is $X_{\max} = -\beta_1/(2\beta_2)$ (a result from calculus). It is a maximum or minimum according to whether β_2 is negative or positive. This is a *nonlinear* combination of coefficients, so the usual method for obtaining standard errors for linear combinations is of no use.

Fieller's Method

Fieller's method tests and constructs confidence intervals for ratios of regression coefficients. The method is illustrated here for X_{\max} .

For a specific value of M , the hypothesis $H_0: X_{\max} = M$ can be re-expressed as $H_0: \beta_1 = -2M\beta_2$. If this hypothesis is correct, the mean for Y above becomes

$$\mu\{Y|X\} = \beta_0 - 2M\beta_2 X + \beta_2 X^2 = \beta_0 + \beta_2 X^*,$$

where $X^* = X^2 - 2MX$ is a new variable, in a simple linear regression model. Applying the extra-sum-of-squares F -test of this as the reduced model against the full model above provides a test of H_0 . A 95% confidence interval is constructed, by trial and error, as the range of M -values that result in test p -values larger than or equal to 0.05. (See Section 4.2.4.)

10.4.6 The Principle of Occam's Razor

The principle of Occam's Razor is that simple models are to be preferred over complicated ones. Named after the 14th-century English philosopher, William of Occam, this principle has guided scientific research ever since its formulation. It has no underlying theoretical or logical basis; rather, it is founded in common sense and successful experience. It is often called the Principle of Parsimony.

In statistical applications, the idea translates into a preference for the more simple of two models that fit data equally well. One should seek a parsimonious model that is as simple as possible and yet adequately explains all that can be explained. Methods for paring down large sets of explanatory variables are discussed in Chapter 12.

10.4.7 Informal Tests in Model Fitting

Tests for hypotheses about regression coefficients—*t*-tests and extra-sum-of-squares *F*-tests—are valuable for two purposes: for formally providing evidence regarding questions of interest in a final model and for exploring models by testing potential terms at the exploratory stage. The attitude toward the *p*-values is somewhat different for these two purposes.

In answering questions of interest, *p*-values are given their formal interpretation. A small *p*-value provides evidence against the null hypothesis and in favor of the alternative. The null hypothesis may be true and the particular sample may have happened by chance to be an unusual one; the *p*-value provides a measure of just how unusual it would be. A large *p*-value means either that the null hypothesis is correct or that the study was not powerful enough to detect a departure from it.

For example, the adequacy of a straight line regression model may be examined through casual testing of an additional quadratic term. The statistical significance of the quadratic term helps to clarify possible curvature. In this usage of testing, some decision—in this case about a suitable model—is made. It should be realized that this decision is sample size dependent; one is more likely to find evidence of curvature in a large data set than in a small one. Nevertheless, the device of informal testing is useful in conjunction with other exploratory tools.

For answering questions of interest, tests should not be overemphasized. Even if the question of interest calls for a test, reporting a confidence interval to indicate the possible sizes of the effects of interest remains important. This is true whether the *p*-value for the test is small or large.

10.5 SUMMARY

Galileo's Study

Galileo's data are used to find a polynomial describing the mean distance as a function of height. The scatterplot shows that the relationship is not a straight line. The coefficient of a height-squared term, when added to the simple linear regression model, significantly differs from zero. R^2 is a useful summary here: the quadratic

regression model explains 99.03% of the variation in the horizontal distances. The large R^2 does not mean that all other terms are insignificant. A test of hypothesis is used to resolve that matter. In fact, when a height-cubed term is added to the model, the p -value for testing whether its coefficient is zero is 0.007, and the value of R^2 increases to 99.94%. This example demonstrates the benefit of R^2 for summarizing a fit and for indicating the degree of relevance of a significant term.

Echolocation by Bats

The goal is to see whether the in-flight energy expended by echolocating bats differs from the energy used by non-echolocating bats of similar body mass. There is no major question involving the birds, but their inclusion helps to clarify a common relationship between energy and body mass. A useful starting point in the analysis is a coded scatterplot of energy versus body mass. It is apparent by inspection that both energy and body mass should be transformed to their logarithms. A plot on this scale (Display 10.4) reveals a straight line relationship but very little additional difference in energy expenditure among the three types of flying vertebrates. Comparing the in-flight energy expenditures of echolocating bats and non-echolocating bats is difficult because the former are all small bats and the latter all big bats. Multiple linear regression permits a comparison after accounting for body mass, but the comparison must be made cautiously, since the ranges of body mass for the two types do not overlap. In light of this limitation, the comparison is made on the basis of an indicator variable in a parallel regression lines model. The data are consistent with the hypothesis that echolocating bats pay no extra energy price for their echolocating skills.

10.6 EXERCISES

Conceptual Exercises

- Galileo's Data.** Why is horizontal distance, rather than height, the response variable?
- Brain Weight.** Display 10.16 shows two possible *influence diagrams* relating the variables in the brain weight study of Section 9.1.2. If brain weight is directly associated with gestation period and litter size, as in part (a) of the figure, then animals that have approximately the same body size but different gestation period and different litter size should have different brain sizes, and scatterplots of brain size versus gestation and litter size individually should show some association. But the scatterplots can also show indirect associations, as in part (b) of the figure. If brain weight, gestation period, and litter size are all driven by body size, they should show mutual associations, whether or not a direct association exists. Can a statistical analysis distinguish between direct and indirect association? Explain how or, if not, why not.
- Brain Weight.** Consider the mammal brain weight data from Section 9.1.2, the model

$$\mu\{lbrain \mid lbody, lgest, llitter\} = \beta_0 + \beta_1 lbody + \beta_2 lgest + \beta_3 llitter,$$

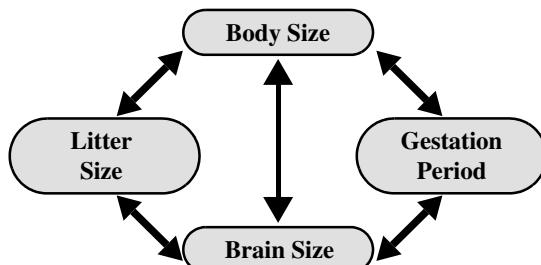
and the hypothesis $H: \beta_2 = 0$ and $\beta_3 = 0$. (a) Why can this not be tested by the two t -tests reported in the standard output? (b) Why can this not be tested by the two t -tests along with an adjustment for multiple comparisons?

DISPLAY 10.16

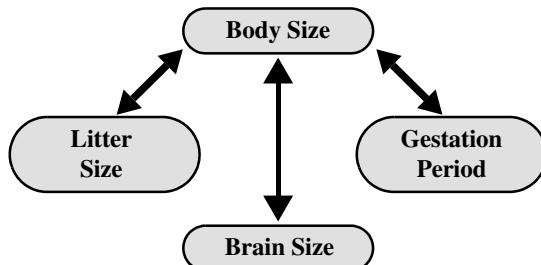
Associations of brain weight with gestation period and litter size: direct or indirect?

(a) DIRECT ASSOCIATION

Animals with about the same body size will have different brain sizes when their litter sizes and gestation periods are different.

**(b) INDIRECT ASSOCIATION**

Animals with about the same body size will have about the same brain size even if their gestation periods and litter sizes are different. Any apparent association between brain size and litter size, say, is a result of their both being related to body size.



4. Stocks and Jocks. Based on data from nine days in June 1994, a multiple regression equation was fit to the Dow Jones Index on the following seven explanatory variables: the high temperature in New York City on the previous day; the low temperature on the previous day; an indicator variable taking on the value 1 if the forecast for the day was sunny and 0 otherwise; an indicator variable taking on the value 1 if the New York Yankees won their baseball game of the previous day and 0 if not; the number of runs the Yankees scored; an indicator variable taking on the value of 1 if the New York Mets won their baseball game of the previous day and 0 if not; and the number of runs the Mets scored. As the chart in Display 10.17 shows, the predicted values of the stock market index were strikingly close to the actual values. R^2 was 89.6%. Why is this unremarkable?

5. Bat Echolocation. Consider these three models:

$$\mu\{lenergy \mid lmass, TYPE\} = \beta_0 + \beta_1 lmass$$

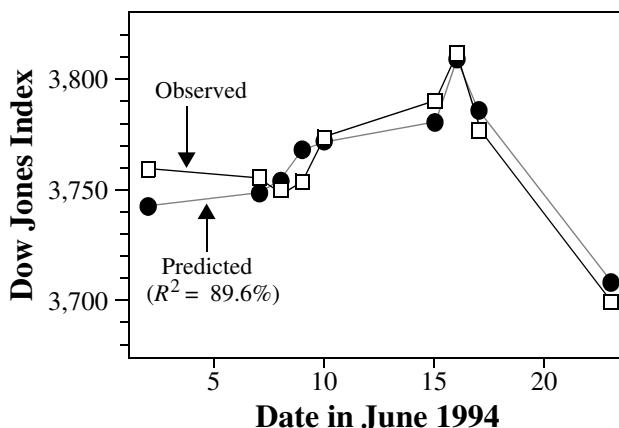
$$\mu\{lenergy \mid lmass, TYPE\} = \beta_0 + \beta_1 lmass + \beta_2 bird + \beta_3 ebat$$

$$\mu\{lenergy \mid lmass, TYPE\} = \beta_0 + \beta_1 lmass + \beta_2 bird + \beta_3 ebat$$

$$+ \beta_4(lmass) \times (bird) + \beta_5(lmass \times ebat).$$

- (a) Explain why they can be described as representing the single line, parallel lines, and separate lines models, respectively. (b) Explain why the second model can be the “reduced model” in one F -test but the “full model” in another.

6. Bat Echolocation. A possible statement in conclusion of the analysis is this: “It is estimated that the median in-flight energy expenditure for echolocating bats is 1.08 times as large as the median in-flight energy expenditure for non-echolocating bats of similar body size.” Referring to Display 10.4, explain why including the phrase, “of similar body size” in this statement is suspect. What alternative wording is available?

DISPLAY 10.17 Actual values and predicted values of Dow Jones Index

7. Life Is a Rocky Road. A regression of the number of crimes committed in a day on volume of ice cream sales in the same day showed that the coefficient of ice cream sales was positive and significantly differed from zero. Which of the following is the most likely explanation? (a) The content of ice cream (probably the sugar) encourages people to commit crimes. (b) Successful criminals celebrate by eating ice cream. (c) A pathological desire for ice cream is triggered in a certain percentage of individuals by certain environmental conditions (such as warm days), and these individuals will stop at nothing to satisfy their craving. (d) Another variable, such as temperature, is associated with both crime and ice cream sales.

8. In the terminology of extra-sums-of-squares F -tests, does the reduced model correspond to the case where the null hypothesis is true? Is the full model the one that corresponds to the alternative hypothesis's being true?

Computational Exercises

9. Crab Claws. Reconsider the data on claw closing force and claw size for three species of crabs, shown in Exercise 7.22. Display 10.18 shows output from the least squares fit to the separate lines model for the regression of log force on log height. The regression model for log force was

$$\begin{aligned}\mu\{\text{lforce} \mid \text{lheight}, \text{SPECIES}\} = & \beta_0 + \beta_1 \text{lheight} + \beta_2 \text{lb} + \beta_3 \text{cp} \\ & + \beta_4 (\text{lheight} \times \text{lb}) + \beta_5 (\text{lheight} \times \text{cp}),\end{aligned}$$

where lheight represents log height, lb represents an indicator variable for the second species, and cp represents an indicator variable for the third species. The sample size was 38.

- (a) How many degrees of freedom are there in the estimate of σ^2 ?
- (b) What is the p -value for the test of the hypothesis that the slope in the regression of log force on log height is the same for species 2 as it is for species 1?
- (c) What is a 95% confidence interval for the amount by which the slope for species 3 exceeds the slope for species 1?

10. Crab Claws. The sum of squared residuals from the fit described in Exercise 9 is 5.99713, based on 32 degrees of freedom. The sum of squared residuals from the fit without the last two terms

DISPLAY 10.18

Least squares output for Exercise 9

Variable	Estimate	SE	t-stat	p-value
Constant	0.5191	1.0000	0.5191	0.6073
<i>lheight</i>	0.4083	0.4868	0.8387	0.4079
<i>lb</i>	-4.2992	1.5283	2.8131	0.0083
<i>cp</i>	-2.4864	1.7606	1.4123	0.1675
<i>lheight</i> × <i>lb</i>	2.5653	0.7354	3.4885	0.0014
<i>lheight</i> × <i>cp</i>	1.6601	0.7889	2.1043	0.0433

DISPLAY 10.19

Regression output data for Exercise 11

Variable	Estimate	SE	t-stat	p-value
Constant	3.775	0.3881	9.7321	<0.0000
<i>lsize</i>	0.0809	0.1131	0.7139	0.2443
<i>days</i>	0.0774	0.1447	0.5346	0.5104

Estimated SD about the regression is 0.8234 on 13 degrees of freedom; $R^2 = 11.41\%$.

is 8.38155, based on 34 degrees of freedom. Form an F -statistic and find the p -value for the test that the slopes are the same for the three species.

11. Butterfly Occurrences. Display 10.19 summarizes results from the regression of the log of the number of butterfly species observed on the log of the size of the reserve and the number of days of observations, from 16 reserves in the Amazon River Basin.

- (a) What is the two-sided p -value for the test of whether size of reserve has any effect on number of species, after accounting for the days of observation? What is the one-sided p -value if the alternative is that size has a positive effect? Does this imply that there is no evidence that the median number of species is related to reserve size? The researchers tended to spend more days searching for butterflies in the larger reserves. How might this affect the interpretation of the results?
- (b) What is a two-sided p -value for the test that the coefficient of *lsize* is 1? (This is simply a computational exercise; there is no obvious reason to conduct this test with these data.)
- (c) What is a 95% confidence interval for the coefficient of *lsize*?
- (d) What proportion of the variation in log number of species remains unexplained by log size and days of observations?

12. Brain Weights. With the data described in Section 9.1.2, construct an extra sum of squares F -test for determining whether gestation period and litter size are associated with brain weight after body weight is accounted for.

13. Bat Echolocation. (a) Fit the parallel regression lines model to duplicate the results in Display 10.6. (b) From these results, what are the estimated intercept and estimated slope for the regression of log energy on log mass for (i) non-echolocating bats, (ii) non-echolocating birds, and (iii) echolocating bats? (c) Refit the model using, instead, the indicator variables *bird* and *nbat*, where *nbat* takes on the value 1 for species of non-echolocating bats and 0 for other species. (d) Based on the results in part (c), what are the estimated intercept and estimated slope for the regression of log energy on log mass for (i) non-echolocating bats, (ii) non-echolocating birds, and (iii) echolocating bats? How

DISPLAY 10.20

Protein in minnow larvae exposed to copper and zinc

Copper (ppm)	Zinc (ppm)	Protein ($\mu\text{g/larva}$)	Copper (ppm)	Zinc (ppm)	Protein ($\mu\text{g/larva}$)
0	0	201	112.5	0	188
0	375	186	112.5	375	172
0	750	173	112.5	750	157
0	1,125	110	112.5	1,125	115
0	1,500	115	112.5	1,500	108
37.5	0	202	150	0	133
37.5	375	161	150	375	125
37.5	750	172	150	750	184
37.5	1,125	138	150	1,125	135
37.5	1,500	133	150	1,500	114
75	0	204			
75	375	165			
75	750	148			
75	1,125	143			
75	1,500	123			

do these compare to the estimates obtained in part (b)? (e) With the results of (c), test whether the lines for the echolocating bats and the non-echolocating birds coincide.

14. Toxic Effects of Copper and Zinc. In a study of the joint toxicity of copper and zinc, researchers randomly allocated 25 beakers containing minnow larvae to receive one of 25 treatment combinations. The treatment levels were all combinations of 5 levels of zinc and 5 levels of copper added to a beaker. Following a four-day exposure, a sample of the minnow larvae were homogenized and analyzed for protein. The results are shown in Display 10.20. (Data from D. A. J. Ryan, J. J. Hubert, J. B. Sprague, and J. Parrott, “A Reduced-Rank Multivariate Regression Approach to Aquatic Joint Toxicity Experiments,” *Biometrics* 48 (1992): 155–62.) Fit a full second-order model for the regression of protein on copper and zinc, and examine the plot of residuals versus fitted values. Repeat after taking the log of protein. Which model is preferable?

15. Kentucky Derby. Reconsider the Kentucky Derby winning times and speeds from Exercise 9.20. Test whether there is any effect of the categorical factor “Track” (with seven categories) on winning speed, after accounting for year. The full model will have *Year* and the categorical factor *Track*; the reduced model will have only *Year*.

16. Galileo’s Data. Use Galileo’s data in Display 10.1 (data file: *case1001*) to perform the following operations.

- (a) Fit the regression of distance on height and height-squared. Obtain the estimates, their standard errors, the estimate of σ^2 , and the variance–covariance matrix of the estimated coefficients.
- (b) Verify that the square roots of the diagonal elements are equal to the standard errors reported with the estimated coefficients.
- (c) Compute the estimated mean distance when the initial height is 500 punti.
- (d) Calculate the standard error for the estimated mean in part (c).
- (e) Use the answer to parts (a) and (d) and the relationship between the variance of the estimated mean and the variance of prediction to obtain the standard error of prediction at an initial height of 500 punti.

- 17. Galileo's Data.** Use Galileo's data in Display 10.1 (data file case1001) to fit the regression of distance on (a) height; (b) height and height²; (c) height, height², and height³; (d) height, height², height³, and height⁴; (e) height, height², height³, height⁴, and height⁵; (f) height, height², height³, height⁴, height⁵, and height⁶. For each part, find R^2 and R_{adj}^2 .
- 18. Corn Yield and Rainfall.** Reconsider the corn yield and rainfall data (Display 9.18; data file ex0915). Fit the regression of yield on rainfall, rainfall-squared, and year. Use the approach of Section 10.4.5 to find the rainfall that maximizes mean yield.
- 19. Meadowfoam.** Carry out a *lack-of-fit F*-test for the regression of number of flowers on light intensity and an indicator variable for time, using the data in Display 9.2 (data file case0901): (a) Fit the regression of *flowers* on *light* and an indicator variable for *time* = 24, and obtain the analysis of variance table. (b) Fit the same regression except with *light* treated as a factor (using 5 indicator variables to distinguish the 6 groups), and with the interaction of these two factors, and obtain the analysis of variance table. (c) Perform an extra-sum-of-squares *F*-test comparing the full model in part (b) to the reduced model in part (a). (*Note:* The full model contains 12 parameters, which is equivalent to the model in which a separate mean exists for each of the 12 groups. No pattern is implied in this model. See Displays 9.7 and 9.8 for help.)
- 20. Calculus Problem.** The least squares problem in multiple linear regression is to find the parameter values that minimize the sum of squared differences between responses and fitted values,

$$\text{SS}(\beta_0, \beta_1, \dots, \beta_p) = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_{1i} - \dots - \beta_p X_{pi})^2.$$

Set the partial derivatives of SS with respect to each of the unknowns equal to zero. Show that the solutions must satisfy the set of *normal equations*, as follows:

$$\begin{aligned}\beta_0 n + \beta_1 \Sigma X_{1i} + \beta_2 \Sigma X_{2i} + \dots + \beta_p \Sigma X_{pi} &= \Sigma Y_i \\ \beta_0 \Sigma X_{1i} + \beta_1 \Sigma X_{1i}^2 + \beta_2 \Sigma X_{1i} X_{2i} + \dots + \beta_p \Sigma X_{1i} X_{pi} &= \Sigma X_{1i} Y_i \\ \beta_0 \Sigma X_{2i} + \beta_1 \Sigma X_{2i} X_{1i} + \beta_2 \Sigma X_{2i}^2 + \dots + \beta_p \Sigma X_{2i} X_{pi} &= \Sigma X_{2i} Y_i \\ &\vdots && \vdots \\ \beta_0 \Sigma X_{pi} + \beta_1 \Sigma X_{pi} X_{1i} + \beta_2 \Sigma X_{pi} X_{2i} + \dots + \beta_p \Sigma X_{pi}^2 &= \Sigma X_{pi} Y_i,\end{aligned}$$

where each Σ indicates summation over all cases ($i = 1, 2, \dots, n$). Show, too, that solutions to the normal equations minimize SS.

- 21. Matrix Algebra Problem.** Let \mathbf{Y} be the $n \times 1$ column vector containing the responses, let \mathbf{X} be the $n \times (p+1)$ array whose first column consists entirely of ones and whose other columns are the explanatory variable values, and let \mathbf{b} be the $(p+1) \times 1$ column containing the resulting parameter estimates. Show that the normal equations in Exercise 20 can be written in the form

$$(\mathbf{X}^T \mathbf{X}) \mathbf{b} = \mathbf{X}^T \mathbf{Y}.$$

Therefore, as long as the matrix inversion is possible, the least squares solution is

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}.$$

When is the inversion possible?

22. Continuing Exercise 21, statistical theory says that the means of the estimates in the vector \mathbf{AY} , where \mathbf{A} is a matrix, are the elements of the vector $\mathbf{A}\mu\{\mathbf{Y}\}$; and the matrix of covariances of these estimates is $\mathbf{ACov}(\mathbf{Y})\mathbf{A}^T$. Use the theory and the model $\mu\{\mathbf{Y}\} = \mathbf{X}\beta$, $\text{Cov}(\mathbf{Y}) = \sigma^2\mathbf{I}$ (where \mathbf{I} is an $n \times n$ identity matrix) to show that the mean in the sampling distributions of the least squares estimate \mathbf{b} is β . Then show that the matrix of covariances is $\text{Cov}\{\mathbf{b}\} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.

23. Speed of Evolution. Refer back to Exercise 9.18. The authors of that study concluded that although the wing size of North American flies was converging rapidly to the same cline as exhibited by the European flies, the means by which the cline is achieved is different in the North American population.

- (a) As evidence that the means of convergence is different, the authors concluded that there was a marked difference between the NA and the EU patterns of the basal length-to-wing size ratios versus latitude (in females). Fit a multiple linear regression that allows for different slopes and different intercepts. In a single F -test, evaluate the evidence against there being a single straight line that describes the cline on both continents. If you conclude there is a difference, is the difference one of slope alone? of intercept alone? or of both?
- (b) Return to the basic question of whether the wing sizes in NA flies have established a cline similar to their EU ancestors. Using the model developed in Exercise 9.18, answer these questions: (i) Is there a nonzero slope to the cline of NA females? (ii) Is there a nonzero slope to the cline of NA males? (iii) Is there a difference between the clines of NA and EU females, and if so, what is its nature? (iv) Repeat (iii) for males.

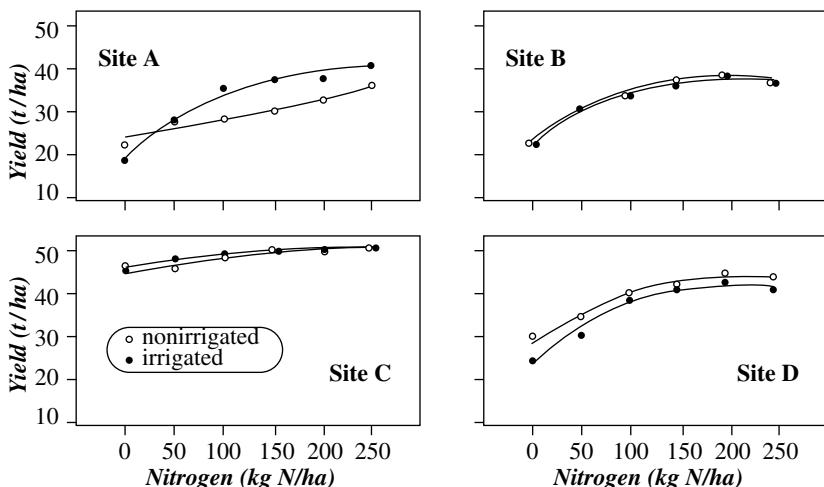
24. Speed of Evolution. (Refer again to Exercise 9.18 and also to Exercise 10.23.) Many software systems allow the user to perform weighted regression, in which different squared residuals from regression receive different weights in deciding which set of parameter estimates provide the smallest sum of squared residuals. If each individual response has an independent estimate of its likely error, the weight given to each residual is usually taken to be the reciprocal of the square of that likely error. The standard error of wing sizes are standard errors of the averages around 2 individual (log) wing sizes. If your software allows for weights, construct a weight variable as the inverse square of the standard errors. Then repeat both parts of Exercise 10.23 using weighted regression. Do the results differ? Why is this preferable to using each fly as a separate case?

25. Potato Yields. Nitrogen and water are important factors influencing potato production. One study of their roles was conducted at sites in the St. John River Valley of New Brunswick. (G. Belanger et al., "Yield Response of Two Potato Cultivars to Supplemental Irrigation and N fertilization in New Brunswick," *American Journal of Potato Research* 77 (2000): 11–21.) Nitrogen fertilizer was applied at six different levels in combination with two water conditions: irrigated or nonirrigated. This design was repeated at four different sites in 1996, with the resulting yields depicted in Display 10.21. Notice that the patterns of responses against nitrogen level are fit reasonably well by quadratic curves.

Each quadratic requires 3 parameters, so a model that would allow for separate quadratic curves for each site-by-irrigation combination would have 24 parameters. (a) Using indicator functions for sites and for irrigation, construct a multiple linear regression model with 23 variables that will allow for completely different quadratic curves. Interpret the parameters in this model, if possible. (b) Describe how you would answer the following questions: (i) Is there evidence that the manner in which the quadratic terms differ by water condition changes from site to site (or is the difference the same at all four sites)? (ii) If the quadratic term differences are the same at all sites, is there strong evidence of a difference by water condition? (iii) If there is no difference between quadratic terms by water or by site, is there evidence of any quadratic term at all? (iv), (v), and (vi): Repeat (i), (ii), and (iii) for the linear terms, if there is no evidence of any quadratic terms. (c) Why are the questions in (b) ordered as they are?

DISPLAY 10.21

Potato yield versus nitrogen at four sites

**Data Problems**

26. Thinning of Ozone Layer. Thinning of the protective layer of ozone surrounding the earth may have catastrophic consequences. A team of University of California scientists estimated that increased solar radiation through the hole in the ozone layer over Antarctica altered processes to such an extent that primary production of phytoplankton was reduced 6 to 12%.

Depletion of the ozone layer allows the most damaging ultraviolet radiation—UVB (280–320 nm)—to reach the earth's surface. An important consequence is the degree to which oceanic phytoplankton production is inhibited by exposure to UVB, both near the ocean surface (where the effect should be slight) and below the surface (where the effect could be considerable).

To measure this relationship, the researchers sampled from the ocean column at various depths at 17 locations around Antarctica during the austral spring of 1990. To account for shifting of the ozone hole's positioning, they constructed a measure of UVB exposure integrated over exposure time. The exposure measurements and the percentages of inhibition of normal phytoplankton production were extracted from their graph to produce Display 10.22. (Data from R. C. Smith et al., "Ozone Depletion: Ultraviolet Radiation and Phytoplankton Biology in Antarctic Waters," *Science* 255 (1992): 952–57.) Does the effect of UVB exposure on the distribution of percentage inhibition differ at the surface and in the deep? How much difference is there? Analyze the data, and write a summary of statistical findings and a section of details documenting those findings. (*Suggestion:* Fit the model with different intercepts and different slopes, even if some terms are not significantly different from zero.)

27. Factors Affecting Extinction. The data in Display 10.23 are measurements on breeding pairs of land-bird species collected from 16 islands around Britain over the course of several decades. For each species, the data set contains an average time of extinction on those islands where it appeared (this is actually the reciprocal of the average of $1/T$, where T is the length of time the species remained on the island, and $1/T$ is taken to be zero if the species did not become extinct on the island); the average number of nesting pairs (the average, over all islands where the birds appeared, of the number of nesting pairs per year); the size of the species (categorized as large or small); and the migratory status of the species (migrant or resident). (Data from S. L. Pimm, H. L. Jones, and

DISPLAY 10.22

First 5 of 17 rows of a data set with exposure to ultraviolet B radiation and percentage inhibition of primary phytoplankton production in Antarctic water

Location	Percent inhibition	UVB exposure	Surface (S) or Deep (D)
1	0.0	0.0000	D
2	1.0	0.0000	D
3	6.0	0.0100	D
4	7.0	0.0150	S
5	7.0	0.0185	S

J. Diamond, “On the Risk of Extinction,” *American Naturalist* 132 (1988): 757–85.) It is expected that species with larger numbers of nesting pairs will tend to remain longer before becoming extinct. Of interest is whether, after accounting for number of nesting pairs, size or migratory status has any effect. There is also some interest in whether the effect of size differs depending on the number of nesting pairs. If any species have unusually small or large extinction times compared to other species with similar values of the explanatory variables, it would be useful to point them out. Analyze the data. Write a summary of statistical findings and a section of details documenting the findings.

28. El Niño and Hurricanes. Shown in Display 10.24 are the first few rows of a data set with the numbers of Atlantic Basin tropical storms and hurricanes for each year from 1950 to 1997. The variable *storm index* is an index of overall intensity of the hurricane season. (It is the average of number of tropical storms, number of hurricanes, the number of days of tropical storms, the number of days of hurricanes, the total number of intense hurricanes, and the number of days they last—when each of these is expressed as a percentage of the average value for that variable. A *storm index* score of 100, therefore, represents, essentially, an average hurricane year.) Also listed are whether the year was a cold, warm, or neutral El Niño year, a constructed numerical variable *temperature* that takes on the values –1, 0, and 1 according to whether the El Niño temperature is cold, neutral, or warm; and a variable indicating whether West Africa was wet or dry that year. It is thought that the warm phase of El Niño suppresses hurricanes while a cold phase encourages them. It is also thought that wet years in West Africa often bring more hurricanes. Analyze the data to describe the effect of El Niño on (a) the number of tropical storms, (b) the number of hurricanes, and (c) the storm index after accounting for the effects of West African wetness and for any time trends, if appropriate. (These data were gathered by William Gray of Colorado State University, and reported on the *USA Today* weather page: www.usatoday.com/weather/whurnum.htm)

29. Wage and Race 1987. Shown in Display 10.25 are the first few rows of a data set from the 1988 March U.S. Current Population Survey. The set contains weekly wages in 1987 (in 1992 dollars) for a sample of 25,437 males between the age of 18 and 70 who worked full-time, their years of education, years of experience, whether they were black, whether they worked in a standard metropolitan statistical area (i.e., in or near a city), and a code for the region in the U.S. where they worked (Northeast, Midwest, South, and West). Analyze the data and write a brief statistical report to see whether and to what extent black males were paid less than nonblack males in the same region and with the same levels of education and experience. Realize that the extent to which blacks were paid differently than nonblacks may depend on region. (Suggestion: Refrain from looking at interactive effects, except for the one implied by the previous sentence.) (These data, from the Current Population Survey (CPS), were discussed in the paper by H. J. Bierens and D. K. Ginther, “Integrated Conditional Moment Testing of Quantile Regression Models,” *Empirical Economics* 26 (2001): 307–24; and made available at the Web site <http://econ.la.psu.edu/~hbierens/MEDIAN.HTM> (April, 2008).)

DISPLAY 10.23

Bird extinction data

Migratory status (resident or migrant)				
Size (large or small)				
Ave. number of nesting pairs				
Ave. extinction time (years)				
Species				
Sparrowhawk	3.030	1.000	L	R
Buzzard	5.464	2.000	L	R
Kestrel	4.098	1.210	L	R
Peregrine	1.681	1.125	L	R
Grey partridge	8.850	5.167	L	R
Quail	1.493	1.000	L	M
Red-legged partridge	7.692	2.750	L	R
Pheasant	3.846	5.630	L	R
Water rail	16.667	3.000	L	R
Corncrake	4.219	4.670	L	M
Moorhen	8.130	4.056	L	R
Coot	5.000	1.000	L	R
Lapwing	7.299	6.960	L	M
Golden plover	1.000	1.670	L	M
Ringed plover	27.027	5.560	L	R
Curlew	3.106	2.830	L	M
Redshank	4.000	4.375	L	M
Snipe	16.129	4.125	L	M
Stock dove	3.484	3.670	L	R
Rock dove	37.037	8.330	L	R
Wood pigeon	7.299	2.750	L	R
Cuckoo	2.525	1.430	L	M
Short-eared owl	4.132	2.000	L	R
Little owl	2.000	2.750	L	R
Magpie	10.000	4.500	L	R
Jackdaw	2.667	7.120	L	R
Carriion crow	4.587	4.580	L	R
Raven	58.824	2.350	L	R
Skylark	32.258	6.870	S	R
Swallow	2.571	3.830	S	M
House martin	2.160	5.000	S	M

Migratory status (resident or migrant)				
Size (large or small)				
Ave. number of nesting pairs				
Ave. extinction time (years)				
Species				
Yellow wagtail	1.000	1.250	S	M
Pied wagtail	2.967	2.270	S	R
Meadow pipit	9.524	5.350	S	R
Wren	11.111	8.700	S	R
Dunnock	7.299	6.100	S	R
Robin	4.000	3.330	S	R
Stonechat	2.381	3.640	S	R
Wheatear	2.611	4.830	S	M
Blackbird	3.257	4.670	S	R
Song thrush	1.701	1.700	S	R
Mistle thrush	1.795	1.330	S	R
Grasshopper warbler	1.198	1.000	S	M
Sedge warbler	3.185	1.900	S	M
Whitethroat	2.273	4.420	S	M
Willow warbler	1.111	1.250	S	M
Chiffchaff	1.000	1.000	S	M
Goldcrest	1.000	1.000	S	R
Spotted flycatcher	1.230	1.000	S	M
Great tit	6.061	2.500	S	R
Blue tit	3.175	1.500	S	R
Yellowhammer	2.000	2.500	S	R
Reed bunting	5.076	5.630	S	R
Chaffinch	1.934	2.370	S	R
Goldfinch	1.493	1.500	S	R
Redpoll	1.000	1.000	S	R
Linnet	5.102	6.500	S	R
House sparrow	3.003	4.500	S	R
Tree sparrow	1.898	2.170	S	R
Starling	41.667	11.620	S	R
Pied flycatcher	1.000	1.000	S	M
Siskin	1.000	1.000	S	R

30. Wages and Race 2011. Display 10.26 is a partial listing of a data set with weekly earnings for 4,952 males between the age of 18 and 70 sampled in the March 2011 Current Population Survey (CPS). These males are a subset who had reported earnings and who responded as having race as either “Only White” or “Only Black.” Also recorded are the region of the country (with four categories: Northeast, Midwest, South, and West), the metropolitan status of the men’s employment (with three categories: Metropolitan, Not Metropolitan, and Not Identified), age, education category (with 16 categories ranging from “Less than first grade” to “Doctorate Degree”), and education code, which is a numerical value that corresponds roughly to increasing levels of education (and so may be useful for plotting). What evidence do the data provide that the distributions of weekly earnings differ in the populations of white and black workers after accounting for the other variables? By how many dollars or by what percent does the White population mean (or median) exceed the Black population mean (or median)? (Data from U.S. Bureau of Labor Statistics and U.S. Bureau of the Census: Current Population Survey, March 2011 http://www.bls.census.gov/cps_ftp.html#cpsbasic; accessed July 25, 2011.)

DISPLAY 10.24 Atlantic Basin hurricane and El Niño data for 1950–1997; partial listing

Year	El Niño	Temperature	West Africa	Storms	Hurricanes	Storm index
1950	cold	-1	1	13	11	243
1951	warm	1	0	10	8	121
1952	neutral	0	1	7	6	97
1953	warm	1	1	14	6	121
...						

DISPLAY 10.25 Data on the first 5 individuals (out of 25,437) in the 1987 wage and race data set

Region	MetropolitanStatus	Exper	Educ	Race	WeeklyEarnings
South	NotMetropolitanArea	8	12	NotBlack	859.71
Midwest	MetropolitanArea	30	12	NotBlack	786.73
West	MetropolitanArea	31	14	NotBlack	1424.5
West	MetropolitanArea	17	16	NotBlack	959.16
West	MetropolitanArea	6	12	NotBlack	154.32

DISPLAY 10.26 First five rows of a data set with weekly earnings (in U.S. dollars) in 2011 for 4,952 males who specified either “White Only” or “Black Only” as their race

Region	MetropolitanStatus	Age	EducCat	EducCode	Race	WeeklyEarnings
West	Not Metropolitan	64	SomeCollegeButNoDegree	40	White	1418.84
Midwest	Metropolitan	51	AssocDegAcadem	42	White	1000.00
South	Metropolitan	25	NinthGrade	35	White	420.00
West	Not Metropolitan	46	MastersDegree	44	White	1980.00
Northeast	Metropolitan	31	BachelorsDegree	43	White	1750.00

31. Who Looks After the Kids? Different bird species have different strategies: Maternal, Paternal, and BiParental care. In 1984 J. Van Rhijn argued in the *Netherlands Journal of Zoology* that parental care was the ancestral condition, going back to the dinosaur predecessors of birds. D. J. Varricchio et al. (*Science*, Vol. 322, Dec. 19, 2008, pp. 1826–28) tested that argument by comparing the relationships between Clutch Volume and adult Body Mass in six different groups: modern maternal-care bird species (*Mat*; $n = 171$), modern paternal-care bird species (*Pat*; $n = 40$), modern biparental-care bird species (*BiP*; $n = 204$), modern maternal-care crocodiles (*Croc*; $n = 19$), non-avian maniraptoran dinosaurs thought to be ancestors of modern birds (*Mani*; $n = 3$), and other non-avian dinosaurs (*Othr*; $n = 6$). The question of interest was which group of modern creatures most closely matches the relationship in the maniraptoran dinosaurs. A partial listing of the data appears in Display 10.27.

Fit a single model, with clutch volume (possibly transformed) as the response, that allows for separate straight lines in each group, using indicator variables for the different groups. For example, by selecting *Mani* as the reference group, you can determine how close any other group is to *Mani* by dropping out the indicator for that group and its product with the body mass variable. Do this for all other groups. (a) Are there differences among the relationships in the groups other than the *Mani*

DISPLAY 10.27

Body mass (in kilograms) and clutch volume (in cubic milimeters) for 443 species of animals in 6 groups (*Mat*: modern maternal-care birds, *Pat*: modern paternal-care birds, *BiP*: modern biparental-care birds, *Croc*: modern maternal-care crocodiles, *Mani*: non-avian maniraptoran dinosaurs, and *Othr*: other non-avian dinosaurs; first 5 of 443 rows

CommonName	Genus	Species	Group	BodyMass	ClutchVolume
Mallard	<i>Anas</i>	<i>platyrhynchos</i>	Mat	1.14E+00	5.42E+05
Greylag Goose	<i>Anser</i>	<i>anser</i>	Mat	3.31E+00	7.97E+05
Mute Swan	<i>Cygnus</i>	<i>olor</i>	Mat	1.07E+01	1.75E+06
Common Eider	<i>Somateria</i>	<i>mollissima</i>	Mat	2.07E+00	4.75E+05
Southern Screamer	<i>Chauna</i>	<i>torquata</i>	BiP	4.40E+00	5.13E+05

DISPLAY 10.28

First five rows of a data set with component test scores in Arithmetic Reasoning, Word Knowledge, Paragraph Comprehension, and Mathematics Knowledge taken in 1981; the AFQT score, which is a linear combination of them; and annual income in 2005 for 12,139 Americans who were selected in the NLSY79 sample, who were available for re-interview in 2006 and who had complete values of variables used in this and related analyses

Subject	Arith	Word	Parag	Math	AFQT	Income2005
2	8	15	6	6	6.841	5,500
6	30	35	15	23	99.393	65,000
7	14	27	8	11	47.412	19,000
8	13	35	12	4	44.022	36,000
9	21	28	10	13	59.683	65,000

group? What if there were none? (b) What do you conclude about which other group is nearest the maniraptoran dinosaur group? (c) Is the model defensible? (d) Comment on the study design.

32. Galton's Height Data. Reconsider the data in Exercise 7.26 on heights of adult children and their parents. Ignore the possible dependence of observations on children from the same family for now to answer the following: (a) What is an equation for predicting a child's adult height from their mother's height, their father's height, and their gender? (b) By how much does the male mean height exceed the female mean height for children whose parents' heights are the same? (c) Find a 95% prediction interval for a female whose father's height is 72 inches and whose mother's height is 64 inches.

33. IQ Score and Income. Display 10.28 is a partial listing of the National Longitudinal Study of Youth (NLSY79) subset (see Exercise 2.22) with annual incomes in 2005 (in U.S. dollars, as recorded in a 2006 interview) and scores on the Word Knowledge, Paragraph Comprehension, Arithmetic Reasoning, and Mathematics Knowledge portions of the Armed Forces Vocational Aptitude Battery (ASVAB) of tests taken in 1981. The AFQT is the Armed Forces Qualifying Test percentile, which is based on a linear combination of these four components and which is sometimes used as a general intelligence test score. A previous exercise had to do with the possible dependence of 2005 income on AFQT score. An interesting question is whether there might be some better linear combination of the four components than AFQT for predicting income. Investigate this by seeing whether the component scores are useful predictors of Income2006 in addition to AFQT. Also see whether

AFQT is a useful predictor in addition to the four test scores. Which test scores seem to be the most important predictors of 2006 income? (Answer with a single p -value conclusion.)

Answers to Conceptual Exercises

1. The initial heights were controlled by Galileo. Distance is the only random quantity.
2. Yes. The extra-sum-of-squares F -test, which compares the model with all three explanatory variables to the model with $lbody$ only, addresses precisely this issue.
3. (a) The test where β_2 and β_3 both equal zero can be cast in terms of comparing the full model ($\beta_0 + \beta_1 lbody + \beta_2 lgest + \beta_3 llitter$) to the reduced model ($\beta_0 + \beta_1 lbody$). The t -test where β_2 is zero, on the other hand, implies a reduced model of $\beta_0 + \beta_1 lbody + \beta_3 llitter$; and the t -test where β_3 is zero implies a reduced model of $\beta_0 + \beta_1 lbody + \beta_2 lgest$. Neither of these reduced models is the same as the one sought, nor can the results from them be combined in any way to give some answer. (b) Same reason. The t -tests consider models where only one parameter is zero, but the model with both β_1 and β_2 equal to zero does not enter the picture. Multiple comparison adjustment does nothing to resolve the fact that these tests are different.
4. The model has nearly as many free parameters (8) as it has observations (9). One should expect a good fit to the data at hand, even if the explanatory variables have little relationship to the response.
5. (a) Each of the models represents the relationship between $lenergy$ and $lmass$ as a straight line, within the groups. The first model says that the intercept and slope—and hence the full line—is the same in all groups. The second model says that the slope is the same in each model while the intercepts are different. The third model allows the slopes and the intercepts to differ among all groups. (b) The second model is a reduced model in a test for equal slopes (with possibly differing intercepts); it is the full model for a test of equal intercepts (given equal slopes).
6. No bats of comparable size could be found in both groups. A more correct wording here might be “after adjustment for body size,” but the underlying difficulty is not avoided.
7. (d).
8. The reduced model takes the null hypothesis to be true. The full model, however, encompasses both the null hypothesis and the alternative hypothesis. Thus, the full model is also correct when the reduced model is correct. Put another way, the full model is thought to be adequate from the start; the reduced model is obtained by imposing the constraints of the null hypothesis on the full model.

Model Checking and Refinement

Multiple regression analysis takes time and care. Dead ends in the pursuit of models are expected and common, especially when many explanatory variables are involved. Frustration and wasted effort can be avoided, however, by going about the analysis in a proper order. In particular, transformations and outliers must be dealt with early on. Although each analysis will be guided by the peculiarities of the particular data and the questions of interest, initial assessment and graphical analysis are usually followed by the fitting of a rich, tentative model and by examination of residual plots. This part of the analysis should suggest whether more investigation into transformation or whether examination of outliers is needed. Some special tools for the latter are provided in this chapter. Once the data analyst has checked that the inferential tools are valid and not seriously influenced by one or two observations, the structure of the model itself can be refined by testing terms to see which should be included.

11.1 CASE STUDIES

11.1.1 Alcohol Metabolism in Men and Women—An Observational Study

Women exhibit a lower tolerance for alcohol and develop alcohol-related liver disease more readily than men. When men and women of the same size and drinking history consume equal amounts of alcohol, the women on average carry a higher concentration of alcohol in their bloodstream. According to a team of Italian researchers, this occurs because alcohol-degrading enzymes in the stomach (where alcohol is partially metabolized before it enters the bloodstream and is eventually metabolized by the liver) are more active in men than in women. The researchers studied the extent to which the activity of the enzyme explained the first-pass alcohol metabolism and the extent to which it explained the differences in first-pass metabolism between women and men. Their data (read from a graph) are listed in Display 11.1. (Data from M. Frezza et al., “High Blood Alcohol Levels in Women,” *New England Journal of Medicine* 322 (1990): 95–99.)

The subjects were 18 women and 14 men, all volunteers living in Trieste. Three of the women and five of the men were categorized as alcoholic. All subjects received ethanol, at a dose of 0.3 grams per kilogram of body weight, orally one day and intravenously another, in randomly determined order. Since the intravenous administration bypasses the stomach, the difference in blood alcohol concentration—the

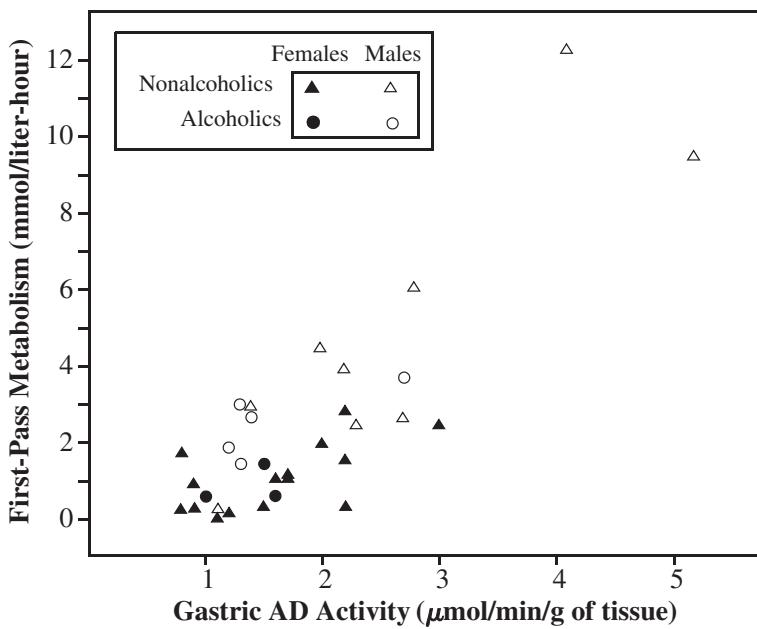
DISPLAY 11.1

First-pass metabolism of alcohol in the stomach (mmol/liter-hour) and gastric alcohol dehydrogenase activity in the stomach ($\mu\text{mol}/\text{min/g}$ of tissue) for 18 women and 14 men

Alcoholic (1 = A, 0 = N)					Alcoholic (1 = A, 0 = N)				
Female (1 = F, 0 = M)					Female (1 = F, 0 = M)				
Gastric activity					Gastric activity				
Metabolism					Metabolism				
Subject					Subject				
1	0.6	1.0	1	1	17	2.5	3.0	1	0
2	0.6	1.6	1	1	18	2.9	2.2	1	0
3	1.5	1.5	1	1	19	1.5	1.3	0	1
4	0.4	2.2	1	0	20	1.9	1.2	0	1
5	0.1	1.1	1	0	21	2.7	1.4	0	1
6	0.2	1.2	1	0	22	3.0	1.3	0	1
7	0.3	0.9	1	0	23	3.7	2.7	0	1
8	0.3	0.8	1	0	24	0.3	1.1	0	0
9	0.4	1.5	1	0	25	2.5	2.3	0	0
10	1.0	0.9	1	0	26	2.7	2.7	0	0
11	1.1	1.6	1	0	27	3.0	1.4	0	0
12	1.2	1.7	1	0	28	4.0	2.2	0	0
13	1.3	1.7	1	0	29	4.5	2.0	0	0
14	1.6	2.2	1	0	30	6.1	2.8	0	0
15	1.8	0.8	1	0	31	9.5	5.2	0	0
16	2.0	2.0	1	0	32	12.3	4.1	0	0

DISPLAY 11.2

First-pass metabolism and gastric alcohol dehydrogenase activity in alcoholic and nonalcoholic men and women



concentration after intravenous administration minus the concentration after oral administration—provides a measure of the “first-pass metabolism” in the stomach. In addition, gastric alcohol dehydrogenase (AD) activity (activity of the key enzyme) was measured in mucus samples taken from the stomach linings. The data are plotted in Display 11.2.

Several questions arise. Do levels of first-pass metabolism differ between men and women? Can the differences be explained by postulating that men have more dehydrogenase activity in their stomachs? Are the answers to these questions complicated by an alcoholism effect?

Statistical Conclusion

The following inferences pertain only to individuals with gastric AD activity levels between 0.8 and 3.0 $\mu\text{mol}/\text{min}/\text{g}$. No reliable model could be determined for values greater than 3.0. There was no evidence from these data that alcoholism was related to first-pass metabolism in any way (p -value = 0.93, from an F -test for significance of alcoholism and its interaction with gastric activity and sex.) Convincing evidence exists that first-pass metabolism was larger for males than for females overall (two-sided p -value = 0.0002, from a rank-sum test) and that gastric AD activity was larger for males than for females (two-sided p -value = 0.07 from a rank-sum test). Males had higher first-pass metabolism than females even after accounting for differences in gastric AD activity (two-sided p -value = 0.0003 from a t -test for

equality of male and female slopes when both intercepts are zero). For a given level of gastric dehydrogenase activity, the mean first-pass alcohol metabolism for men is estimated to be 2.20 times as large as the mean first-pass alcohol metabolism for women (approximate 95% confidence interval from 1.37 to 3.04).

Scope of Inference

Because the subjects were volunteers, no inference to a larger population is justified. The inference that men and women do have different first-pass metabolism is greatly strengthened, however, by the existence of a physical explanation for the difference. The conclusions about the relationship between first-pass metabolism, gastric AD dehydrogenase activity, and sex are restricted to individuals whose gastric AD activity is less than 3. The sparseness of data for individuals with greater gastric AD activity levels prevents any resolution of the answers in the wider range.

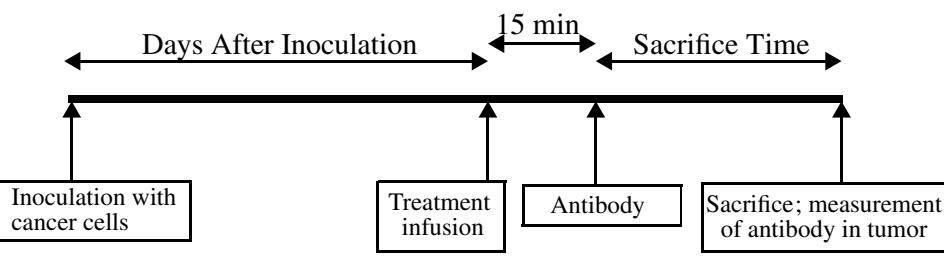
11.1.2 The Blood–Brain Barrier—A Controlled Experiment

The human brain is protected from bacteria and toxins, which course through the bloodstream, by a single layer of cells called the *blood–brain barrier*. This barrier normally allows only a few substances, including some medications, to reach the brain. Because chemicals used to treat brain cancer have such large molecular size, they cannot pass through the barrier to attack tumor cells. At the Oregon Health Sciences University, Dr. E. A. Neuwelt developed a method of disrupting the barrier by infusing a solution of concentrated sugars.

As a test of the disruption mechanism, researchers conducted a study on rats, which possess a similar barrier. (Data from P. Barnett et al., “Differential Permeability and Quantitative MR Imaging of a Human Lung Carcinoma Brain Xenograft in the Nude Rat,” *American Journal of Pathology* 146(2) (1995): 436–49.) The rats were inoculated with human lung cancer cells to induce brain tumors. After 9 to 11 days they were infused with either the barrier disruption (BD) solution or, as a control, a normal saline (NS) solution. Fifteen minutes later, the rats received a standard dose of the therapeutic antibody L6-F(ab')₂. After a set time they were sacrificed, and the amounts of antibody in the brain tumor and in normal tissue were measured. The time line for the experiment is shown in Display 11.3. Measurements for the 34 rats are listed in Display 11.4.

DISPLAY 11.3

Time line for blood–brain barrier disruption experiment



DISPLAY 11.4

Response variable, design variables (explanatory variables associated with the assignment of experimental units to groups), and several covariates (explanatory variables not associated with the assignment) for 34 rats in the blood-brain barrier disruption experiment

Case	Response variable		Design variables		Covariates				
	Brain tumor count (per gm)		Sacrifice time (hours)						
	Liver count (per gm)		Treatment		Days post inoculation	Tumor weight (10^{-4} grams)	Weight loss (grams)	Initial weight (grams)	Sex
1	41081/ 1456164		0.5	BD	10	F	239	5.9	221
2	44286/ 1602171		0.5	BD	10	F	225	4.0	246
3	102926/ 1601936		0.5	BD	10	F	224	-4.9	61
4	25927/ 1776411		0.5	BD	10	F	184	9.8	168
5	42643/ 1351184		0.5	BD	10	F	250	6.0	164
6	31342/ 1790863		0.5	NS	10	F	196	7.7	260
7	22815/ 1633386		0.5	NS	10	F	200	0.5	27
8	16629/ 1618757		0.5	NS	10	F	273	4.0	308
9	22315/ 1567602		0.5	NS	10	F	216	2.8	93
10	77961/ 1060057		3	BD	10	F	267	2.6	73
11	73178/ 715581		3	BD	10	F	263	1.1	25
12	76167/ 620145		3	BD	10	F	228	0.0	133
13	123730/ 1068423		3	BD	9	F	261	3.4	203
14	25569/ 721436		3	NS	9	F	253	5.9	159
15	33803/ 1019352		3	NS	10	F	234	0.1	264
16	24512/ 667785		3	NS	10	F	238	0.8	34
17	50545/ 961097		3	NS	9	F	230	7.0	146
18	50690/ 1220677		3	NS	10	F	207	1.5	212
19	84616/ 48815		24	BD	10	F	254	3.9	155
20	55153/ 16885		24	BD	10	M	256	-4.7	190
21	48829/ 22395		24	BD	10	M	247	-2.8	101
22	89454/ 83504		24	BD	11	F	198	4.2	214
23	37928/ 20323		24	NS	10	F	237	2.5	224
24	12816/ 15985		24	NS	10	M	293	3.1	151
25	23734/ 25895		24	NS	10	M	288	9.7	285
26	31097/ 33224		24	NS	11	F	236	5.9	380
27	35395/ 4142		72	BD	11	F	251	4.1	39
28	18270/ 2364		72	BD	10	F	223	4.0	153
29	5625/ 1979		72	BD	10	M	298	12.8	164
30	7497/ 1659		72	BD	10	M	260	7.3	364
31	6250/ 928		72	NS	10	M	272	11.0	484
32	11519/ 2423		72	NS	11	F	226	2.2	168
33	3184/ 1608		72	NS	10	M	249	-4.4	191
34	1334/ 3242		72	NS	10	F	240	6.7	159

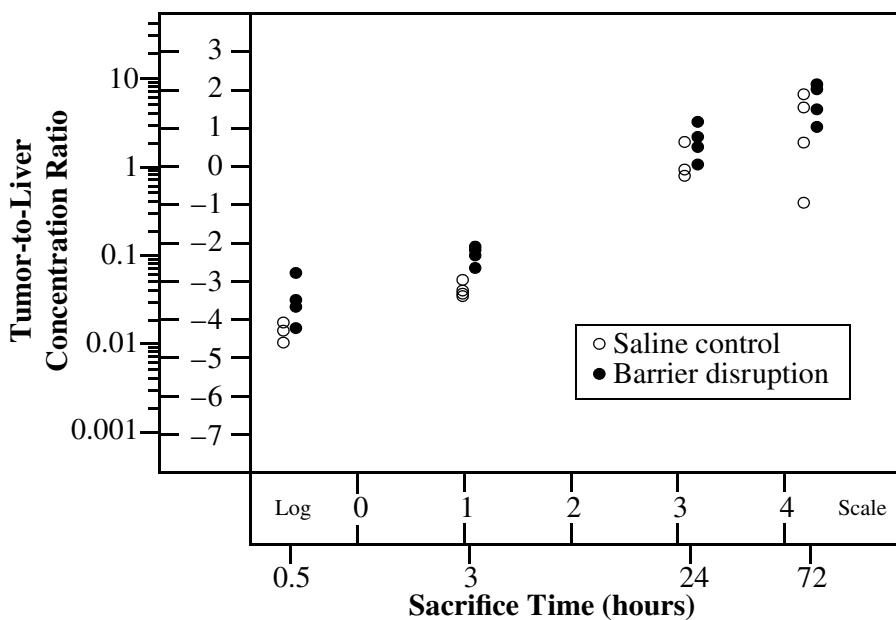
Since the amount of the antibody in normal tissue indicates how much of it the rat actually received, a key measure of the effectiveness of transmission across the blood-brain barrier is the ratio of the antibody concentration in the brain tumor

to the antibody concentration in normal tissue outside of the brain. The brain tumor concentration divided by the liver concentration is a measure of the amount of the antibody that reached the brain relative to the amount of it that reached other parts of the body. This is the response variable: both the numerator and denominator of this ratio are listed in Display 11.4. The explanatory variables in the table comprise two categories: *design variables* are those that describe manipulation by the researcher; *covariates* are those measuring characteristics of the subjects that were not controllable by the researcher.

Was the antibody concentration in the tumor increased by the use of the blood-brain barrier disruption infusion? If so, by how much? Do the answers to these two questions depend on the length of time after the infusion (from 1/2 to 72 hours)? What is the effect of treatment on antibody concentration after weight loss, total tumor weight, and other covariates are accounted for? A coded scatterplot relating to the major questions is shown in Display 11.5.

DISPLAY 11.5

Log-log scatterplot of the ratio of antibody concentration in brain tumor to antibody concentration in liver versus sacrifice time, for 17 rats given the barrier disruption infusion and for 17 rats given a saline (control) infusion


Statistical Conclusion

The median antibody concentration in the tumor (relative to that in the liver) was estimated to be 2.22 times as much for rats receiving the barrier disruption infusion than for those receiving the control infusion (95% confidence interval, from 1.56 to

3.15 times as much). This multiplicative effect appears to be constant between 1/2 and 72 hours after the infusion (the p -value for a test of interaction between treatment and sacrifice time is 0.92, from an F -test on 3 and 26 degrees of freedom).

Scope of Inference

One hitch in this study is that randomization was not used to assign rats to treatment groups. This oversight raises the possibility that the estimated relationships might be related to confounding variables over which the experimenter exercised no control. Including the measured covariates in the model helps alleviate some concern, and the results appear not to have been affected by these potential confounding variables. Nevertheless, causal implications can only be justified on the tenuous assumption that the assignment method used was as effect-neutral as a random assignment would have been.

11.2 RESIDUAL PLOTS

Faced with analyzing data sets like those involved in the blood–brain barrier and alcohol metabolism studies, a researcher must seek good-fitting models for answering the questions of interest, bearing in mind the model assumptions required for least squares tools, the robustness of the tools against violations of the assumptions, and the sensitivity of these tools to outliers. Since model-building efforts are wasted if the analyst fails to detect problems with nonconstant variance and outliers early on, it is wise to postpone detailed model fitting until after outliers and transformation have been thoroughly considered.

Much can be resolved from initial scatterplots and inspection of the data, but it is almost always worthwhile to obtain the finer picture provided by a residual plot. Creating this plot involves fitting some model in order to get residuals. On the basis of the scatterplots, the analyst can choose some tentative model or models and conduct residual analysis on these, recognizing that further modeling will follow.

Selecting a Tentative Model

A tentative model is selected with three general objectives in mind: The model should contain parameters whose values answer the questions of interest in a straightforward manner; it should include potentially confounding variables; and it should include features that capture important relationships found in the initial graphical analysis.

It is disadvantageous to start with either too many or too few explanatory variables in the tentative model. With too few, outliers may appear simply because of omitted relationships. With too many (lots of interactions and quadratic terms, for example), the analyst risks overfitting the data—causing real outliers to be explained away by complex, but meaningless, structural relationships. Overfitting becomes less of a problem when the sample sizes are substantially larger than the number of model parameters.

For large sample sizes, therefore, the initial tentative model for residual analysis can err on the side of being rich, including potential model terms that may not be retained in the end. For small sample sizes, several tentative models may be needed for residual analysis; and the data analyst must guard against including terms whose significance hinges on one or two observations. As evident in the strategy for data analysis laid out in Display 9.9, the process of trying a model and plotting residuals is often repeated until a suitable inferential model is determined.

Example—Preliminary Steps in the Analysis of the Blood-Brain Barrier Data

The coded scatterplot in Display 11.5 is a good starting point for the analysis. Apparently, the disruption solution does allow more antibody to reach the brain than the control solution does; this effect is about the same for all sacrifice times (time between antibody treatment and sacrifice); an increasing proportion of antibody reaches the brain with increasing time after infusion; and this increasing relationship appears to be slightly nonlinear. A matrix of scatterplots and a correlation matrix (an array showing the sample correlation coefficients for all possible pairs of variables), which are not shown here, indicate further that the covariates—days after inoculation, initial weight, and sex of the rat—are associated with the response. These covariates are also related to the treatment given. (Recall that randomization was not used.) In particular, rats treated at longer days after inoculation were also assigned to the longer sacrifice times. Furthermore, all male rats were assigned to the longer sacrifice times.

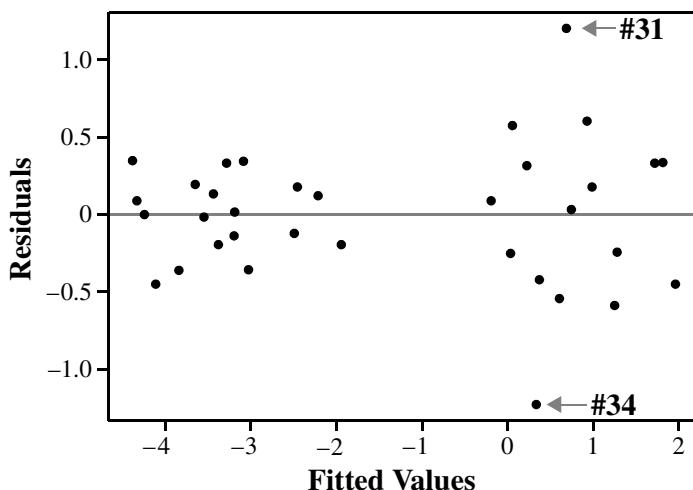
This initial investigation suggests the following tentative regression model (using the shorthand model specification of Section 9.3.5):

$$\begin{aligned}\mu\{\text{antibody} | \text{SAC, TREAT, DAYS, FEM, weight, loss, tumor}\} \\ = \text{SAC} + \text{TREAT} + (\text{SAC} \times \text{TREAT}) + \text{DAYS} + \text{FEM} + \text{weight} + \text{loss} + \text{tumor},\end{aligned}$$

where *antibody* is the logarithm of the ratio of antibody in the brain tumor to that in the liver. *SAC* is the sacrifice time factor with four levels; *TREAT* is treatment, with two levels; *DAYS* is days after inoculation, with three levels; and *FEM* is sex, with two levels. *Weight*, *loss*, and *tumor* are the initial weight, weight loss, and tumor weight variables. Display 11.5 shows a strong linear effect of log sacrifice time on the response, but some additional curvature may be present as well. To avoid mismodeling the effect of sacrifice time at the start, it is treated as a factor with four levels. Similarly, the coded scatterplot suggests that the difference between the two treatments may be greater for the shorter sacrifice times than for the longer ones. Consequently, the sacrifice time by treatment interaction terms are included in the tentative model. Although more terms may be added to this model later, it captures the most prominent features of the scatterplot. Display 11.6 shows the plot of residuals versus the fitted values from the regression model. (*Note:* Even if prior experience or initial inspection had not led the researchers to consider the logarithms of the response, the coded scatterplot and residual plot would have revealed that the variability increases with increasing response, leading them to the same consideration.)

DISPLAY 11.6

Scatterplot of residuals versus fitted values from the fit of the logged response on a rich model for explanatory variables—brain-barrier data



The residual plot in Display 11.6 exemplifies the ambiguity that can arise with small data sets. Is there a funnel-shaped pattern, or is the apparent funnel only due to a few outliers? The usual course of action consists of three steps:

1. Examine the outliers for recording error or contamination.
2. Check whether a standard transformation resolves the problem.
3. If neither of these steps works, examine the outliers more carefully to see whether they influence the conclusions (following the strategy suggested in Section 11.3).

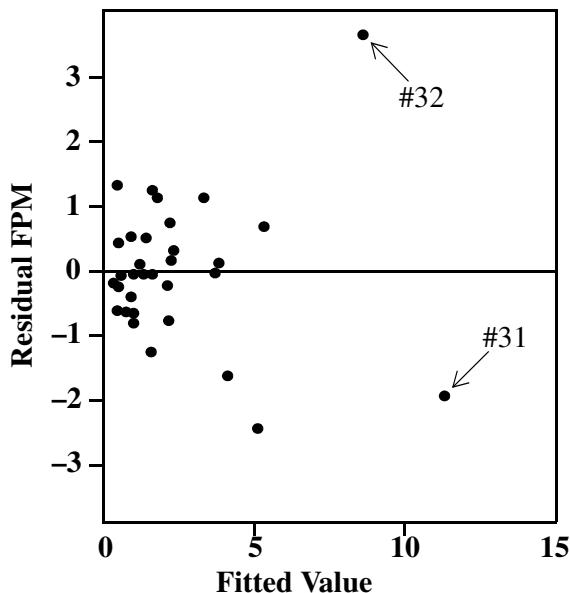
The residual plot is based on a response that has already been transformed into its logarithm. A reciprocal transformation corrects more pronounced funnel-shaped patterns than does the log. Here, however, it does not help, and there is no suggestion of a recording error. Consequently, the analyst must proceed with further model fitting, paying careful attention to the roles of observations 31 and 34.

Example—Preliminary Steps in the Analysis of the Alcohol Metabolism Data

Refer to the coded scatterplot in Display 11.2. The next step is to examine a residual plot for outliers and to assess the need for transformation. The plot in Display 11.7 is a residual plot from the regression of first-pass metabolism on gastric AD activity (*gast*), an indicator variable for females (*fem*), an indicator for alcoholics (*alco*), and the interaction terms *gast* \times *fem*, *fem* \times *alco*, *gast* \times *alco*, and *gast* \times *fem* \times *alco*. (The last term is a *three-factor interaction* term, formed as the product of three explanatory variables.)

DISPLAY 11.7

Residual plot from the regression of first-pass metabolism on gastric activity, sex indicator, alcoholism indicator, and all two- and three-factor interactions



The plot draws attention to two observations: one that has a considerably larger residual than the rest and one that has a fitted value quite a bit larger than the rest. These are cases 31 and 32, and they appear in the coded scatterplot of Display 11.2 in the upper right-hand corner, separated from the rest of the points. There appears to be a downward trend in the residual plot, excluding cases 31 and 32. This could reflect a model that is heavily influenced by one or two observations and consequently does not fit the bulk of the observations well.

11.3 A STRATEGY FOR DEALING WITH INFLUENTIAL OBSERVATIONS

Least squares regression analysis is not resistant to outliers. One or two observations can strongly influence the analysis, to the point where the answers to the questions of interest change when these isolated cases are excluded. Although any influential observation that comes from a population other than the one under investigation should be removed, removing an observation simply because it is influential is not justified. In any circumstance, it is unwise to state conclusions that hinge on one or two data points. Such a statistical study should be considered extremely fragile.

There are two approaches for dealing with excessively influential observations in regression analysis. One is to use a robust and resistant regression procedure. The other is to use least squares but to examine outliers and influence closely to

see whether the suspect observations are indeed influential, why they are influential (this can help dictate the subsequent course of action), and whether they provide some interesting extra information about the process under study. Using a robust and resistant regression procedure is particularly useful if past experience indicates that the response distribution tends to have long tails and that outliers are an expected nuisance. When influential observations are discovered in the course of a regression analysis, however, *The Statistical Sleuth* recommends closer examination. This often clarifies the problems and sometimes reveals additional, unexpected information.

Assessment of Whether Observations Are Influential

The strategy for assessing influence involves temporarily removing suspected influential observations to see whether the answers to the questions of interest change. Does the evidence from a test change from slight evidence to convincing evidence? Does the decision to include a term in the model change? Does an important estimate change by a practically relevant amount? If not, the observation is not influential and the analysis can proceed as usual. If so, further action must be taken.

What to Do About Influential Observations

If an observation is influential and it substantially differs from the remaining data in its explanatory variable values, perhaps it is being accorded too much weight in fitting a model over a sparsely represented region of the explanatory variables. For example, given 30 observations for X between 0 and 5 and a single value of X at 15, it is unrealistic to try to model the regression of Y on X over the entire range of 0 to 15. The observations in the sparsely represented region can be removed, the model fit on the remaining data points, and conclusions stated only for the restricted range of explanatory variables. This restriction should be stated explicitly, with the added comment that more data would be needed to model the regression accurately over the broader range of explanatory variables.

If an influential observation is not particularly unusual in its explanatory variable values, and if no definitive explanation for its unique behavior can be found, omitting it cannot be justified. More data are needed to answer the questions of interest. As a last resort, the results can be reported with and without the influential observation.

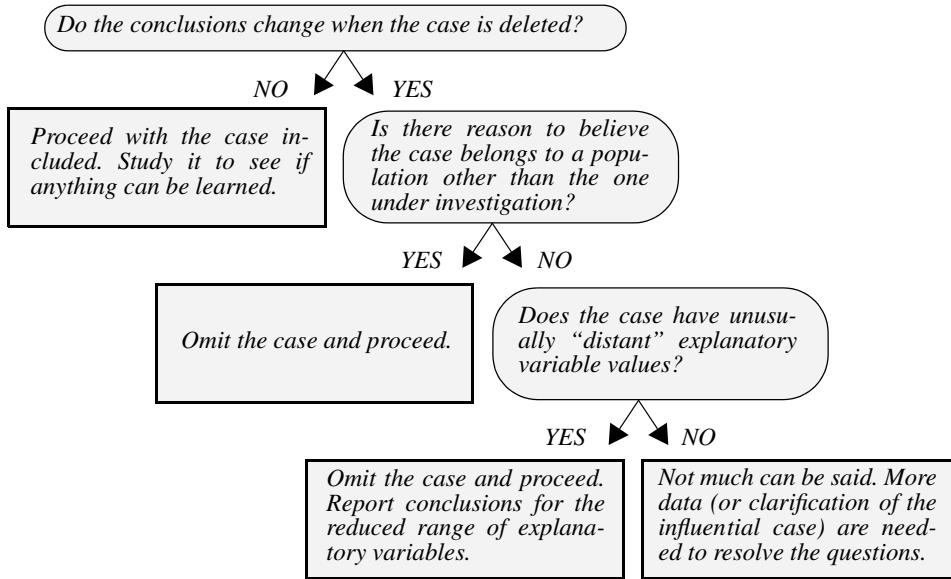
A complete strategy is summarized in Display 11.8. Although this is not stated explicitly in the strategy, the analyst should first check any unusual observation for recording accuracy and for alternative explanations of its unusualness before proceeding to this statistical approach.

Example—Alcohol Metabolism Study

The tentative model is fit with and without cases 31 and 32, to examine which aspects of the fit change and by how much. The resulting changes in estimates, standard errors, and p -values are shown in Display 11.9.

DISPLAY 11.8

A strategy for dealing with suspected influential cases

**DISPLAY 11.9**

Regression parameter estimates, standard errors, and *p*-values from the regression of first-pass metabolism on gastric activity, an indicator for female, an indicator for alcoholic, and all second- and third-order interactions: first with all observations and then with all observations except 31 and 32

Variable	All 32 observations			Cases 31 and 32 removed		
	Estimate	Standard error	Two-sided <i>p</i> -value	Estimate	Standard error	Two-sided <i>p</i> -value
Constant	-1.660	1.000	0.11	-0.680	1.309	0.61
Gastric activity (G)	2.514	0.343	<0.0001	1.921	0.608	0.0045
Female (F)	1.466	1.333	0.28	0.486	1.467	0.74
Alcoholic (A)	2.552	1.946	0.20	1.572	1.812	0.40
G×F	-1.673	0.620	0.013	-1.081	0.721	0.15
F×A	-2.252	4.394	0.61	-1.272	3.467	0.72
G×A	-1.459	1.053	0.18	-0.866	0.963	0.38
G×F×A	1.199	2.998	0.69	0.606	2.316	0.80

A striking consequence of the exclusion of cases 31 and 32 is the drop in significance of the interaction of gastric activity and sex from a *p*-value of 0.013 to one of 0.15. The reason for this change is evident from the coded scatterplot in Display 11.2. Ignoring the effect of alcoholism, imagine the slope in the regression of first-pass metabolism on gastric AD activity for males and females separately.

The slope for males is substantially greater with cases 31 and 32 included, than with them excluded.

What does this mean? Maybe the estimated slope with those cases is accurate and males have a larger slope than females; but alternatively, perhaps the relationship is not a straight line for gastric activity greater than 3. It is difficult to know how to model the relationship in that region. Proceeding from the guideline that it is unwise to state conclusions that hinge on one or two data points, the prudent action is to exclude cases 31 and 32, and to restrict the model building and conclusions to the restricted range of gastric AD activity less than 3.

11.4 CASE-INFLUENCE STATISTICS

Case-influence statistics are numerical measures associated with the individual influence of each observation (each case). When provided by a statistical computer program, they are useful for two reasons: they can help identify influential observations that may not be revealed graphically; and they partition the overall influence of an observation into what is unusual about its explanatory variable values and what is unusual about its response relative to the fitted model. This partition may be useful in following the strategy suggested in Display 11.8 for dealing with cases of suspect influence.

11.4.1 Leverages for Flagging Cases with Unusual Explanatory Variable Values

The *leverage* of a case is a measure of the distance between its explanatory variable values and the average of the explanatory variable values in the entire data set. As illustrated in the alcohol metabolism study, cases with high leverage may exert strong influence on the results of model fitting.

When only a single explanatory variable X is involved, the leverage of the i th case is

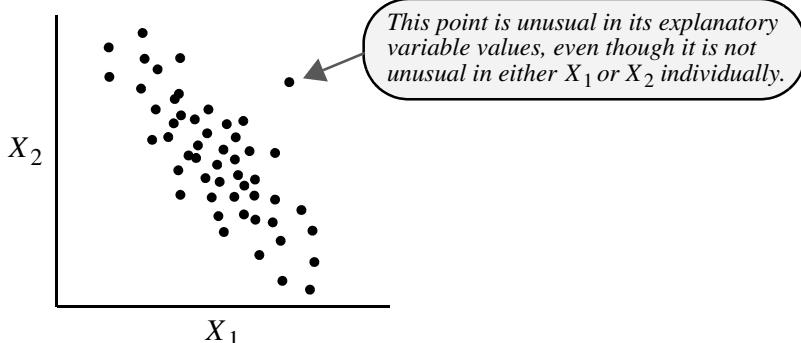
$$h_i = \frac{1}{(n-1)} \left[\frac{X_i - \bar{X}}{s_X} \right]^2 + \frac{1}{n} \quad \text{or} \quad \frac{(X_i - \bar{X})^2}{\sum(X - \bar{X})^2} + \frac{1}{n},$$

where s_X is the sample standard deviation of X . Aside from the parts involving n , the first formulation shows that the leverage is a distance of X_i from \bar{X} , in units of standard deviations; and the second formulation shows that the leverage is the proportion of the total sum of squares of the explanatory variable contributed by the i th case.

When two or more explanatory variables are involved, the leverage can only be expressed in matrix notation. The interpretation is a straightforward extension of the one-variable case, however. The leverage for case i is a measure of the distance of case i from the average (in a multivariable sense). The one aspect of leverage in the multiple regression setting that is not an obvious extension of the preceding formulas is that the distance from the average is relative to the joint spread of the explanatory variables. Display 11.10 illustrates a problem in which one case is

DISPLAY 11.10

An illustration of what is meant by “far from the average” of multiple explanatory variables when they are correlated



relatively far from the two-dimensional scatter of the other cases, although it is not particularly unusual in either of the single dimensions.

It is important to identify this case, since it stands alone. Because it would be largely responsible for dictating the location of the estimated regression surface in the surrounding region, the point has a high potential for influence. The multiple regression version of leverage would be large for such an observation.

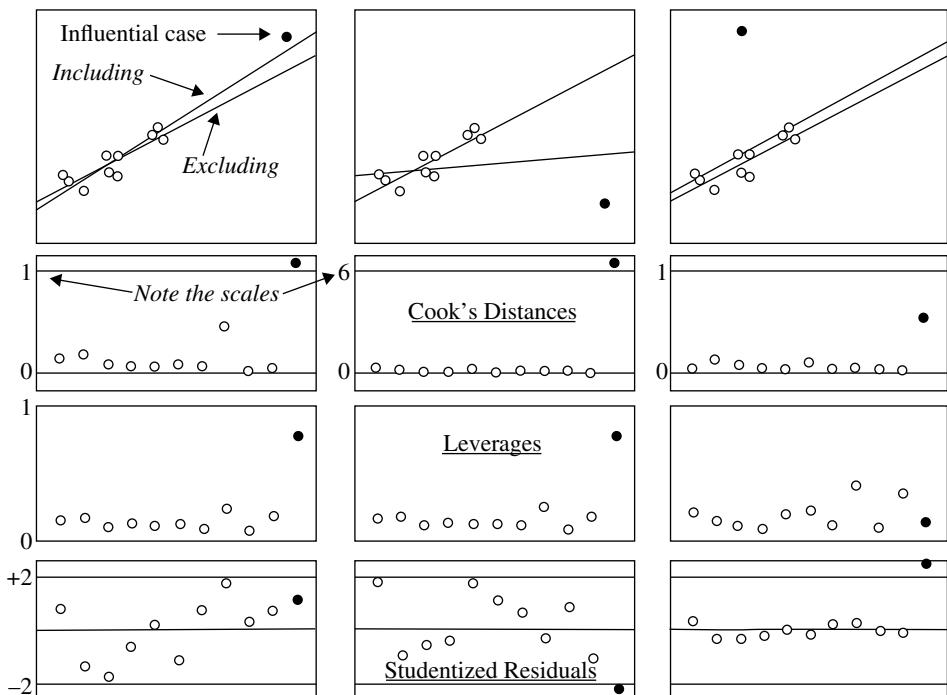
Cases with Large Leverage

Leverages are greater than $1/n$ and less than 1 for each observation, and the average of all the h_i 's in a data set is always p/n , where p is the number of regression coefficients. Leverages also depend on what model is being entertained. The standard deviation of the residual for the i th case is related to its leverage:

$$\text{SD}(\text{Residual}_i) = \sigma \sqrt{(1 - h_i)}.$$

A case with large leverage has a residual with low variability. Because its explanatory variable values are so unusual, it dictates the location of the estimated regression over the whole region in its vicinity; no other points in the region share the responsibility. Because its residual must be small, this case acts like a magnet on the estimated regression surface. If, however, its response falls close to the regression surface (as determined by the remaining observations alone), it is not necessarily influential. Therefore, while a large leverage does not necessarily indicate that the case is influential, it does imply that the case has a high *potential* for influence.

Cases with large leverage can easily be overlooked in scatterplots and other graphical displays. Display 11.10 shows how a case with high leverage can be visible in a two-dimensional scatterplot, even though it would not be exceptional in either of the component one-dimensional histograms. Similarly, a case with high leverage may be visible in a three-dimensional plot but not in any of the component two-dimensional scatterplots. Therefore, examining the numerical measures of leverage

DISPLAY 11.11 Case influence statistics for three sample situations

A. High leverage and mild departure changes the slope so that the residual is small. Cook's Distance identifies the offending case.

B. High leverage and huge departure drastically pulls the line away from all observations. Cook's Distance identifies the case.

C. Low leverage does not allow the large departure to alter the slope, so it ends up with a big residual. Cook's Distance shows a mild problem.

by themselves is important. The leverage measure h_i would identify the case in Display 11.10 because its distance takes into account the correlations among the variables.

It is difficult to say how large a value of h_i is sufficiently large to warrant further attention. Since the average of the h_i 's is p/n , some statisticians (and statistical computer programs) use twice the average— $2p/n$ —as a lower cutoff for flagging cases that have a high potential for excessive influence. This formulation is somewhat arbitrary. The main point is to use the leverage measure in conjunction with the other case influence statistics to get some overall assessment of influence. Display 11.11 will show one way to display the leverages in conjunction with other case influence statistics.

Leverage Computations

Most statistical computer programs will carry out the calculations for leverage as part of their regression routine. If not, the calculations may be made by using the

formula

$$h_i = \left[\frac{\text{SE}(\text{fit}_i)}{\hat{\sigma}} \right]^2.$$

The values for $\text{SE}(\text{fit}_i)$ are often available. If not, they may be found with the method in Section 10.2.3.

11.4.2 Studentized Residuals for Flagging Outliers

A *studentized residual* is a residual divided by its estimated standard deviation. Some residuals naturally have less variation than others because of their leverages. Therefore, the usual residual plot may not direct attention to cases whose residuals lie much farther from zero than expected. The studentized residuals,

$$\text{studres}_i = \frac{\text{res}_i}{\hat{\sigma} \sqrt{1 - h_i}},$$

put all residuals on a common scale: numbers of standard deviations. Since roughly 95% of normally distributed values fall within two standard deviations of their mean, it is common to investigate observations whose studentized residuals are smaller than -2 or larger than 2 . Of course, it is not unusual to find roughly 5% of observations outside this range, and 5% can be a sizeable number if the sample size is large. (Technically, this is called the *internally studentized residual*; see Note 1 at the end of Section 11.4.4.) As with leverages, studentized residual calculations are typically performed by the statistical computer program.

Example—Observation #31 from the First-Pass Metabolism Study

For the fit of the regression of metabolism on gastric activity, sex indicator, and their interaction, the observed value for the 31st observation is 9.5 and its fitted value is 11.0024. This leaves the residual of $\text{res}_{31} = (9.5 - 11.0024) = -1.5024$. From computer calculations, its leverage is $h_{31} = 0.5355$. Notice that the rough lower cutoff for flagging large influence, $2p/n$, is $2 \times 4/32 = 0.25$; according to that rule, case 31 has a high potential for influence. The estimated standard deviation of the residual is $1.20730(1 - 0.5355)^{1/2} = 0.8228$. The studentized residual is -1.8260 . This is moderately far from zero, but not alarmingly so.

11.4.3 Cook's Distances for Flagging Influential Cases

Cook's Distance is a case statistic that measures *overall* influence—the effect that omitting a case has on the estimated regression coefficients. For case i , Cook's Distance can be represented as

$$D_i = \sum_{j=1}^n \frac{(\hat{Y}_{j(i)} - \hat{Y}_j)^2}{p\hat{\sigma}^2},$$

where \hat{Y}_j is the j th fitted value in a fit using all the cases; $\hat{Y}_{j(i)}$ is the j th fitted value in a fit that excludes case i from the data set; p is the number of regression

coefficients; and $\hat{\sigma}^2$ is the estimated variance from the fit, based on all observations. The numerator measures how much the fitted values (and, therefore, the parameter estimates) change when the i th case is deleted. The denominator scales this difference in a useful way. A case that is influential, in terms of the least squares estimated coefficients changing when it is deleted, will have a large value of Cook's Distance.

An equivalent expression for Cook's Distance is

$$D_i = \frac{1}{p} (\text{studres}_i)^2 \left(\frac{h_i}{1 - h_i} \right).$$

This alternate expression is useful for computing, since it does not require that the i th case actually be deleted. More importantly, it shows that an influential case—with a large Cook's Distance—is influential because it has a large studentized residual, a large leverage, or both.

Cook's Distance for case 31 of the alcohol metabolism data is 0.9610. Some statisticians use a rough guideline that a value of D_i close to or larger than 1 indicates a large influence. Realize, however, that Cook's Distance measures the influence on all the regression coefficients. By identifying potentially problematic cases, the data analyst can directly refit the model with and without the indicated observations to see whether the answers to the important questions of interest change. The effect of the case removal on a single coefficient of interest may be much more or much less dramatic than what is indicated by the general measure D_i . Nevertheless, Cook's Distance is very useful for calling attention to potentially problematic cases.

11.4.4 A Strategy for Using Case Influence Statistics

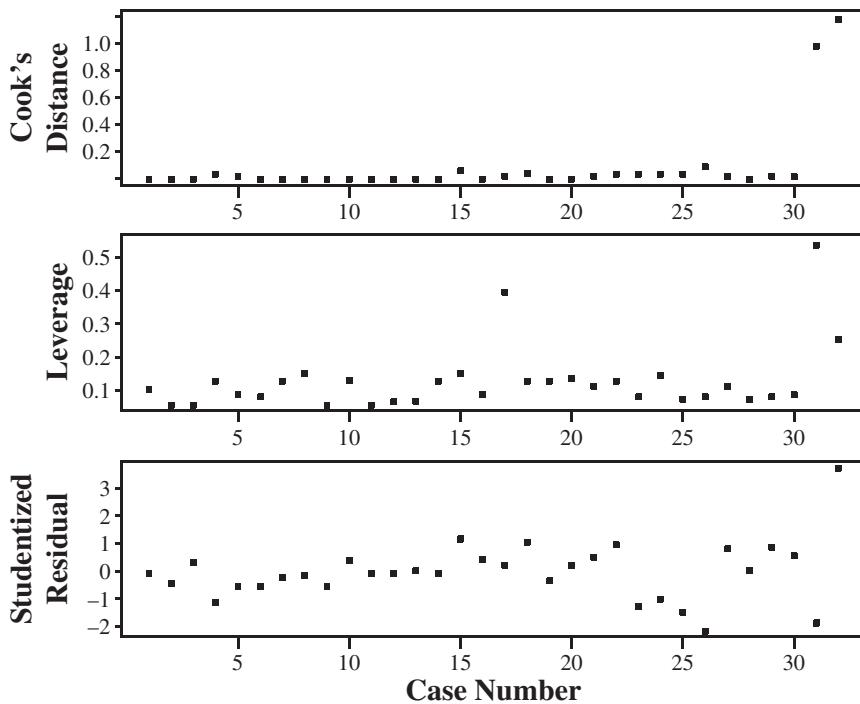
Inspecting graphical displays and scatterplots may alert the analyst to the need for case influence investigations. In addition, the presence of statistically significant, complex effects that make little sense may indicate problems with influence. But when the residual plot from fitting a good inferential model fails to suggest any problems, there is generally no need to examine case influence statistics at all.

The suggested strategy in Display 11.8 requires some assessment of influence, of the unusualness of a case's explanatory variable values, and of the degree to which a case is an outlier. Therefore, the trio of case influence measures— D_i , h_i , and studres_i (or similar trios described later)—are examined jointly. Since values of these measures exist for every observation in the data set, the examination of a full list may be difficult. Some computer programs use rough guidelines (such as $D_i > 1$, $h_i > 2p/n$, or $|\text{studres}_i| > 2$) to flag the relevant cases. Ideally, a graphical display of the case statistics, like Display 11.12 for the alcohol metabolism data, makes case influence analysis convenient.

In the top plot, the values of Cook's Distance for cases 31 and 32 are obviously substantially larger than the rest, indicating what was previously observed: these two observations are influential. In addition, they both have high leverages, as one would expect from their gastric AD activity values' being so much larger than the rest. One other case—number 17—has a large leverage. This is the female with

DISPLAY 11.12

Case-influence statistics for the fit of first-pass metabolism on gastric activity, sex indicator, and their interaction



the largest gastric AD activity value. No evidence indicates that this observation is influential, however, so it need not be studied further.

Notes About Case Influence Statistics

1. *Externally studentized residuals.* A potential problem with the internally studentized residual is that the estimate of σ from the fit with all observations may be tainted by case i if i is indeed an outlier. The *externally studentized residual* is

$$\text{studres}_i^* = \frac{\text{res}_i}{\hat{\sigma}_{(i)} \sqrt{1 - h_i}},$$

where $\hat{\sigma}_{(i)}$ is the estimate of the standard deviation about the regression line from the fit that excludes case i .

2. *External studentization in measures of influence.* If external studentization makes sense for residuals, it also makes sense for the measure of influence. Although it is not widely used, an externally studentized version of Cook's Distance, D_i^* , can be obtained by using $\hat{\sigma}_{(i)}$ in place of $\hat{\sigma}$ in the definition—or, equivalently, by using studres_i^* in place of studres_i in the computing version. A measure that is widely used is DFFITS $_i$, which is $(pD_i^*)^{1/2}$. It measures influence in the

same way as the externally studentized version of Cook's Distance, but on a slightly different scale. The decision about which of these measures to use is based largely on which is available in the statistical computer program.

3. *These case statistics address one-at-a-time influence.* If two or more cases are jointly influential and, in particular, if they are together in an unusual region of the explanatory variables, then Cook's Distance and the leverage measure may fail to detect their joint influence and joint leverage. The user must be on guard for this situation. It is always possible to remove a pair of observations to investigate their joint influence directly.

11.5 REFINING THE MODEL

11.5.1 Testing Terms

The key assumption for the correctness of the statistical conclusions is that the terms in the model sufficiently describe the regression of the response on the explanatory variables. It is misleading to leave out important explanatory variables. On the other hand it is also important to use Occam's Razor (Section 10.4.6) to trim away nonessential terms. Therefore, after transformations and outliers have been resolved, the analysis focuses on finding a simple, good-fitting model.

Example—Alcohol Metabolism Study

Once the decision is made to set cases 31 and 32 aside, the analyst must refit the model and examine the new residual plot. No additional problem is indicated, so model refinement may begin.

Since alcoholism is not of primary concern, and since so few alcoholics are included in the data set, it would be helpful if this variable could be ignored. One approach in the investigation is to test whether all the terms involving alcoholism in the tentative model can be dropped, using an extra-sum-of-squares F -test. There are four such terms: $alco$, $gast \times alco$, $fem \times alco$, and $gast \times fem \times alco$. The F -statistic is 0.21. By comparison to an F -distribution on 4 and 22 degrees of freedom, the p -value is 0.93, so the data are consistent with there being no alcoholism effect.

With alcoholism left out, the resulting model for consideration is

$$\mu\{metabolism | gast, fem\} = \beta_0 + \beta_1 gast + \beta_2 fem + \beta_3 (gast \times fem),$$

the separate regression lines model. The estimates of regression parameters, their standard errors, and their p -values appear in Display 11.13.

It may appear that sex plays no significant role, because the parameters β_2 (the amount by which the women's intercept exceeds the men's intercept) and β_3 (the amount by which the women's slope exceeds the men's slope), have nonsignificant p -values individually. This is deceiving. A sex difference is distinctly noticeable on the scatterplot (Display 11.2). The p -values suggest, however, that there is no need for different slopes if different intercepts are included in the model, and no need for different intercepts if different slopes are included.

DISPLAY 11.13

Least squares estimates for the regression of first-pass metabolism on gastric AD activity, sex, and their interaction (excluding cases 31 and 32)

Variable	Estimate	Standard error	t-statistic	Two-sided p-value
Constant	0.070	0.802	0.087	0.93
<i>gast</i>	1.565	0.407	3.843	0.0007
<i>fem</i>	-0.267	0.993	-0.269	0.79
<i>gast</i> × <i>fem</i>	-0.728	0.539	-1.351	0.19

In Display 11.2, the intercepts for both men and women are near zero. Since, in addition, a zero intercept makes sense in this application (because there is no first-pass metabolism if there is no activity of the enzyme), it is useful to force both lines to go through the origin, by dropping the constant term and the female indicator variable:

$$\mu\{\text{metabolism} \mid \text{gast}, \text{fem}\} = \beta_1 \text{gast} + \beta_2 (\text{gast} \times \text{fem}).$$

In this model, first-pass metabolism is directly proportional to gastric activity, but the constant of proportionality differs for men and for women. The *F*-statistic comparing this model to the one whose fit is summarized in Display 11.13 is 0.06 with 2 and 26 degrees of freedom, so the smaller model is adequate.

The fit to the new model appears in Display 11.14. In this model, both terms are essential, so this is accepted as the final version for inference. The conclusions stated in the summary of statistical findings are based on this fit. Notice in particular that, for any level of gastric AD activity (in the range of 0.8 to 3.0), the mean first-pass metabolism for males divided by the mean first-pass metabolism for females is $\beta_1 / (\beta_1 + \beta_2)$, which is estimated as 2.20. Thus, the mean for males is estimated to be 2.20 times the mean for females, even after accounting for gastric dehydrogenase activity.

DISPLAY 11.14

Results for mean metabolism being proportional to gastric activity

Variable	Estimate	Standard error	t-statistic	Two-sided p-value
<i>gast</i>	1.5989	0.1249	12.800	<0.0001
<i>gast</i> × <i>fem</i>	-0.8732	0.1740	-5.019	<0.0001

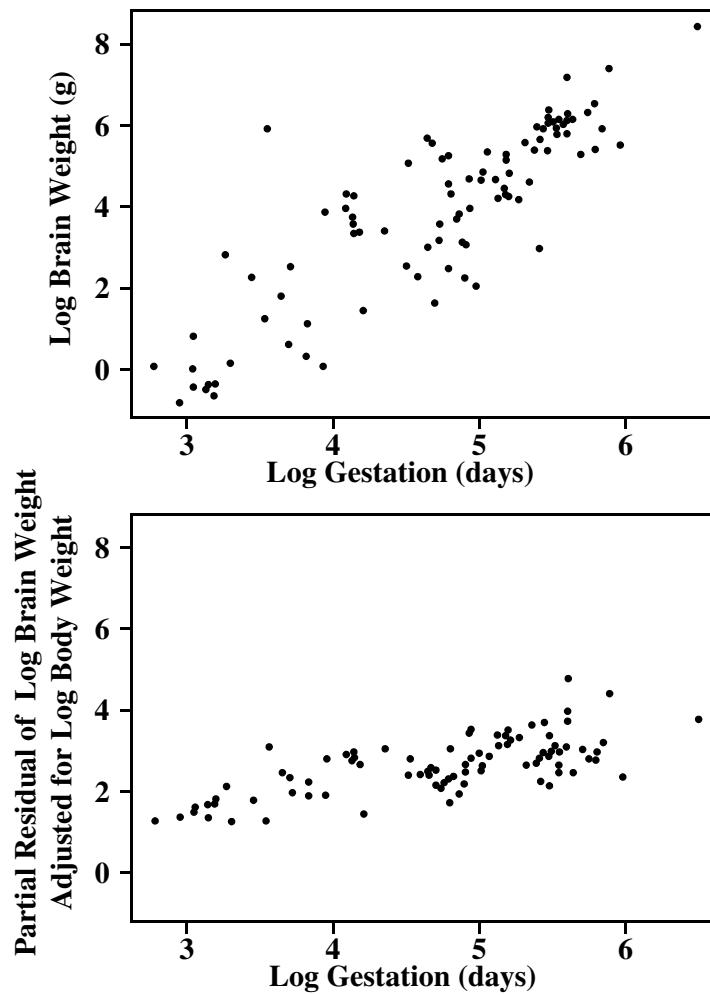
Residual SD = 0.8518; d.f. = 28

11.5.2 Partial Residual Plots

A scatterplot exhibits only the marginal association of two variables, which may depend heavily on their mutual association with a third variable. The question

DISPLAY 11.15

Scatterplot of log brain weight versus log gestation length, and scatterplot of the partial residuals of log brain weight (adjusted for log body weight) versus log gestation length—mammal brain weight data from Section 9.1.2



of interest may be better addressed by a plot showing the association of the two variables, after getting the effect of the third variable out of the way.

The top scatterplot in Display 11.15 shows log brain weight versus log gestation length for the 96 species of mammals examined in the study discussed in Section 9.1.2. An important question of interest is whether an association exists between brain weight and gestation, after body weight is accounted for. The multiple regression coefficient addresses this directly, but a graphical display in conjunction with it would be helpful.

Suppose that

$$\mu\{lbrain \mid lbody, lgest\} = \beta_0 + \beta_1 lbody + f(lgest),$$

where $lbrain$, $lbody$, and $lgest$ represent the logarithms of brain weight, body weight, and gestation length, and where the last term is some unspecified function of $lgest$. The purpose of a *partial residual plot* is to explore the nature of the function $f(lgest)$, including whether it can be adequately approximated by a single linear term $\beta_2 lgest$ and whether the estimated relationship might be affected by one or several influential cases.

In this setup,

$$f(lgest) = \mu\{lbrain \mid lbody, lgest\} - (\beta_0 + \beta_1 lbody),$$

so it would be useful to plot $lbrain - (\beta_0 + \beta_1 lbody)$ versus $lgest$ to visually explore the function $f(lgest)$. Since the β 's are unknown, however, this is impossible. They can be replaced by estimates, but not just any estimates will work. It would not be appropriate to estimate them by the regression of $lbrain$ on $lbody$ alone, since the necessary coefficients β_0 and β_1 are those in the regression of $lbrain$ on $lbody$ and $lgest$. On the other hand, since $f(lgest)$ is unspecified, it is unclear how to include it in the model.

Partial Residuals

The idea behind a partial residual plot is to approximate $f(lgest)$ with the linear function $\beta_2 lgest$. This may be a crude approximation, but it is often good enough to estimate the correct β_0 and β_1 for plotting purposes. The following steps are used to draw a *partial residual plot* of $lbrain$ versus $lgest$, adjusting for $lbody$:

1. Obtain the estimated coefficients in the linear regression of $lbrain$ on $lbody$ and $lgest$: $\mu\{lbrain \mid lbody, lgest\} = \hat{\beta}_0 + \hat{\beta}_1 lbody + \hat{\beta}_2 lgest$.
2. Compute the *partial residuals* as $pres = lbrain - \hat{\beta}_0 - \hat{\beta}_1 lbody$.
3. Plot the partial residuals versus $lgest$.

A partial residual plot is shown in the lower scatterplot of Display 11.15. After getting the effect of log body weight out of the way, less of an association remains between log brain weight and log gestation, but some association does persist and a linear term should be adequate to model it.

When the effects of many other explanatory variables need to be “subtracted,” an easier calculation formula is available. Steps 1 and 2 of the method just described can be replaced with these:

1. Obtain the residuals, res , from the fit to the linear regression of $lbrain$ on $lbody$ and $lgest$ (include all explanatory variables—those whose effects are to be subtracted, and the one that is to be plotted).
2. Compute the partial residuals as $pres = res + \hat{\beta}_2 lgest$ (the residuals from the fit with all explanatory variables plus the estimated component associated with the effect of the explanatory variable under question).

The meaning of the partial residuals is clearer in the first version, but the calculation is often more straightforward with the second. Because of this calculating formula, partial residual plots are sometimes referred to as *component plus residuals plots*.

Notes About Partial Residuals

When Should Partial Residual Plots Be Used? Partial residuals are primarily useful when analytical interest centers on one explanatory variable whose effect is expected to be small relative to the effects of others. They are also useful when uncertainty exists about a particular explanatory variable that needs to be modeled carefully or when the underlying explanation for why an observation is influential on the estimate of a single coefficient needs to be understood.

Augmented Partial Residuals. Rather than using $\beta_2 lgest$ as an approximation to $f(lgest)$, some statisticians prefer to use $\beta_2 lgest + \beta_3 lgest^2$. Here, partial residuals are obtained just as in the preceding algorithms, except that $lgest^2$ is also included as an explanatory variable in step 1. (In step 2 of the component-plus-residual version, $pres = res + \hat{\beta}_2 lgest + \hat{\beta}_3 lgest^2$.) If they are equally convenient to use, the augmented partial residuals are preferred. In many cases, however, the difference between the partial residual and the augmented partial residual is slight.

Example—Blood-Brain Barrier

The residual plot in Display 11.6 indicated some potential outliers, but further investigation does not show that these points are influential in determining the structure of the model or in answering the questions of interest (see Exercise 11.18).

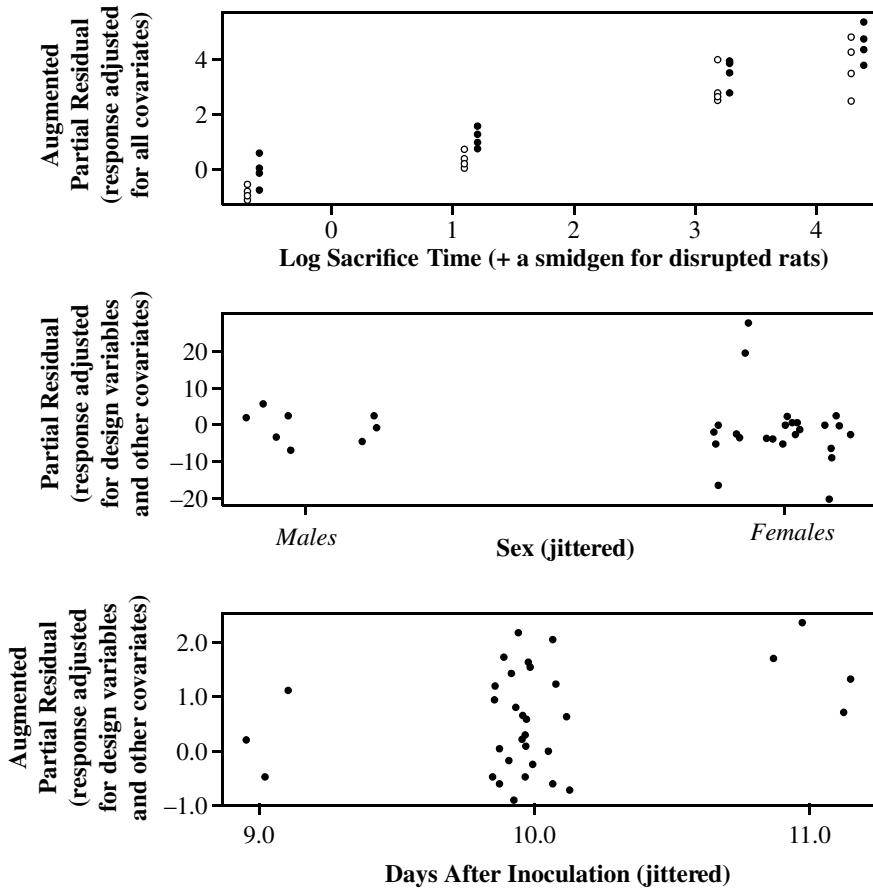
The key explanatory variables are the indicator variable for whether the rat received the disruption infusion or the control infusion, the length of time after infusion that the rat was sacrificed, and the interaction of these. The additional covariates should be given a chance to be included in the model, for two reasons. First (and most importantly), since randomization was not used, it behooves the researchers to demonstrate that the differences in treatment effects cannot be explained by differences in the types of rats that received the various treatments. Second, even if randomization had been used, including important covariates can yield higher resolution. If the covariates have some additional association with the response, smaller standard errors and more powerful tests should result from their inclusion.

Among the covariates, sex and days after inoculation are associated with both the response and the design variables. To some extent the effects of these variables are confounded, since their effects on the response cannot be separated. On the other hand, the effects of the design variables can be examined after the covariates are accounted for, and the effects of the covariates can be examined after the design variables are accounted for. This is shown graphically in the partial residual plots of Display 11.16.

The top scatterplot indicates that the relationship between the response and the design variables (sacrifice time and treatment) is much the same when the

DISPLAY 11.16

Some partial residual plots for the blood–brain barrier data, with log of antibody concentration ratio (brain tumor-to-liver) as response



effects of the covariates are included as when they are ignored (Display 11.5). The lower two plots show that, after the effects of the design variables are accounted for, little evidence exists of a sex effect, although slight visual evidence exists of a days-after-inoculation effect.

This conclusion is further investigated through model fitting. A search through possible models that contain covariates shows that sex and days after inoculation (treated as a factor) are the only ones associated with the response. When the design variables are included as well, three conclusions are supported:

1. The covariates are not significant when the design variables are also included in the model.
2. The design variables *are* significant when the covariates are also included in the model.

DISPLAY 11.17

Results from the regression of log ratio of antibody concentration (brain tumor-to-liver) on sacrifice time (treated as a factor) and treatment

Variable	Estimate	Standard error	t-statistic	Two-sided p-value
Constant	-4.302	0.205	-21.01	<0.0001
Indicator for time = 3	1.134	0.252	4.50	0.0001
Indicator for time = 24	4.257	0.259	16.43	<0.0001
Indicator for time = 72	5.154	0.259	19.89	<0.0001
Indicator for treatment = BD	0.797	0.183	4.35	0.0002

3. The conclusions regarding the design variables depend very little on whether the covariates are in the model.

These results suggest that the conclusions can be based satisfactorily on the model without the covariates.

Since the effect of log sacrifice time is not linear (and since the addition of a quadratic term does not remedy the lack-of-fit), sacrifice time is treated as a factor with four levels. Therefore, the final model used to estimate the treatment effect has the following terms: *TIME* + *TREAT*. The estimates and standard errors are shown in Display 11.17. The coefficient of the indicator variable for the blood–brain barrier disruption treatment is 0.797. So, expressed in accordance with the interpretation for log-transformed responses, the median ratio of antibody concentration in the brain tumor to antibody concentration in the liver is estimated to be $\exp(0.797) = 2.22$ times greater for the blood–brain barrier diffusion treatment than for the control.

11.6 RELATED ISSUES

11.6.1 Weighted Regression for Certain Types of Nonconstant Variance

Although nonconstant variance can sometimes be corrected by a transformation of the response, in many situations it cannot. If enough information is known about the form of the nonconstant variance, the method of *weighted least squares* may be used.

The *weighted regression* model, written here with two explanatory variables, is

$$\begin{aligned}\mu\{Y_i \mid X_{1i}, X_{2i}\} &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \\ \text{Var}\{Y_i \mid X_{1i}, X_{2i}\} &= \sigma^2/w_i,\end{aligned}$$

where the w_i 's are known constants called *weights* (because cases with larger w_i 's have smaller variances and should be weighted more in the analysis).

This model arises in at least three practical situations:

1. *Responses are estimates; SEs are available.* Sometimes the response values are measurements whose estimated standard deviations, $\text{SE}(Y_i)$, are available. In the preceding model, the w_i 's are taken to be $1/[\text{SE}(Y_i)]^2$; that is, the responses with smaller standard errors should receive more weight.
2. *Responses are averages; only the sample sizes are known.* If the responses are averages from samples of different sizes and if the ordinary regression model applies for the individual observations (the ones going into the average), then the weighted regression model applies to the averages, with weights equal to the sample sizes. The averages based on larger samples are given more weight.
3. *Variance is proportional to X.* Sometimes, while the regression of a response on an explanatory variable is a straight line, the variance increases with increases in the explanatory variable. Although a log transformation of the response might correct the nonconstant variance, it would induce a nonlinear relationship. A weighted regression model, with $w_i = 1/X_i$ (or possibly $w_i = 1/X_i^2$) may be preferable.

The weighted regression model can be estimated by *weighted least squares* within the standard regression procedure in most statistical computing programs. The estimated regression coefficients are chosen to minimize the weighted sum of squared residuals (see Exercise 21 for the calculus). It is necessary for the user to specify the response, the explanatory variables, and the weights.

11.6.2 The Delta Method

When, as in the alcohol metabolism study, there is a quantity of interest that is a nonlinear function of model parameters, calculating a standard error for the estimate of the quantity requires advanced methods. One such method—the *delta method*—requires some calculus and is therefore presented as an optional topic.

Taking the alcohol metabolism study's example, suppose interest centers on the parameter $\theta = \beta_1/(\beta_1 + \beta_2)$, where estimates of β_1 and β_2 are available. Substituting the estimates into the equation for θ produces an estimate for θ . Two inputs are required for calculating its standard error: (1) the variance–covariance matrix of the β -estimates, which should be available from the computer, and (2) the partial derivatives of θ with respect to each of the β 's. Display 11.18 illustrates how these pieces combine to produce the standard error for this θ .

11.6.3 Measurement Errors in Explanatory Variables

Sometimes a theoretical model specifies that the mean response depends on certain explanatory variables that cannot be measured directly. This is called the *errors-in-variables problem*. If, for example, a study is examining the relationship between blood cholesterol (Y) and the dietary intake of polyunsaturated fat (X), and if the intake of polyunsaturated fat is estimated from a questionnaire individuals supply on what they eat in a typical week, then the questionnaire results will not measure X precisely.

DISPLAY 11.18

The delta method for calculating the standard error of a nonlinear function of parameter estimates, applied to the alcohol metabolism study

- ① Express the parameter of interest, θ , as a function of parameters estimated by computer analysis.

$$\theta = \frac{\beta_1}{\beta_1 + \beta_2}$$

- ② Obtain β -estimates and their variance–covariance matrix from the computer.

	<u>Estimate</u>	<u>Variance–Covariance Matrix</u>	
β_1	1.5989	0.01561	-0.01561
β_2	-0.8732	-0.01561	0.03027
		$\hat{\text{Var}}(\hat{\beta}_1)$	$\hat{\text{Cov}}(\hat{\beta}_1, \hat{\beta}_2)$
			$\hat{\text{Var}}(\hat{\beta}_2)$

- ③ Calculate and estimate the partial derivatives of θ with respect to the β 's.

$$\theta_1 = \frac{\partial \theta}{\partial \beta_1} = \frac{\beta_2}{(\beta_1 + \beta_2)^2} \approx \frac{-0.8732}{(1.5989 - 0.8732)^2} = -1.6581$$

$$\theta_2 = \frac{\partial \theta}{\partial \beta_2} = \frac{-\beta_1}{(\beta_1 + \beta_2)^2} \approx \frac{-1.5989}{(1.5989 - 0.8732)^2} = -3.036$$

- ④ The estimate of θ has approximate variance

$$\text{Var}(\hat{\theta}) = \sum_i \theta_i^2 \text{Var}(\hat{\beta}_i) + 2 \sum_{i < j} \theta_i \theta_j \text{Cov}(\hat{\beta}_i, \hat{\beta}_j)$$

$$\approx (-1.6581)^2(0.01561) + (-3.036)^2(0.03027) + 2(-1.6581)(-3.036)(-0.01561) = 0.1648$$

- ⑤ The (approximate) standard error of the estimate is the square root ... = $\sqrt{0.1648} = 0.4060$

If the purpose of the regression is prediction, and if the prediction of future responses is to be based on measured explanatory variables that have the same kind of measurement errors as the ones used to estimate the regression, then measurement errors in explanatory variables do not present a problem. The model of interest in this case is the one for the regression of the response on the measured explanatory variables. The regression on the exact explanatory variables is not relevant.

For other purposes, however, measurement errors in explanatory variables present a problem. The least squares estimates based on the imprecisely measured explanatory variables are biased estimates of the coefficients in the regression of the response on the precisely measured explanatory variables. If there is a single explanatory variable, the slope estimate tends to be closer to zero than it should be. If there are many explanatory variables, some of which are measured with error, the individual coefficients (including the coefficients of variables that are free of measurement error) may be over- or underestimated, depending on the correlation structure of the explanatory variables.

Alternatives to least squares that account for measurement errors in explanatory variables do exist, but they are sophisticated and require extra information about the measurement errors (for example, that replicate measurements are

available on a subset of subjects). The advice of a professional statistician may be required for this problem.

11.7 SUMMARY

This chapter focuses on the first three steps—exploring the data, fitting a model, and checking the model—of the strategy for data analysis using statistical models (Display 9.9). After initial exploration, the analyst fits a tentative model and uses the residuals to check on the need for transformation or outlier action. The exploration continues by entertaining various models and testing terms within the model. There is not necessarily any “best” or “final” model. Often different questions of interest are addressed through different models. In the blood–brain barrier data set, for example, one of the questions of interest is investigated via interaction terms. When these are found to be insignificant, they are dropped in favor of a different model for making inferences about the treatment effect.

Alcohol Metabolism Study

Several questions are asked of these data. For two-sample questions like “Does the first-pass metabolism of alcohol differ between males and females?” the rank-sum test is used to avoid considering the outliers. To investigate whether metabolism differs between males and females once the effect of gastric AD activity has been accounted for, a regression model for metabolism as a function of gastric AD activity and an indicator variable for sex should be used. Finding a model is made difficult by the need to consider quite a few terms on the basis of a small data set, as well as by the influence of two observations with atypical values of the explanatory variable. Since the data are not capable of resolving a regression relationship for gastric AD activities larger than 3, the two extreme cases are set aside and the subsequent analysis and conclusions are restricted to the reduced range. Within that reduced range (and ignoring the effect of alcoholism), either of two models can be used: the parallel regression lines model or the regression model forced through the origin with different slopes for males and females. The latter is selected for inference primarily because it seems somewhat more likely to apply in the broader range (and because it results in a smaller estimate of residual variation).

Blood–Brain Barrier Study

The lack of random assignment of the rats to the treatment groups is a major concern. In partial compensation for this design flaw, tentative regression models include potentially confounding covariates—measured but uncontrolled characteristics of the experimental units. Although some covariate differences are detected among the treatment groups, the regression analysis is able to clarify the treatment effects after accounting for the covariates. Without randomization, however, the causal conclusions remain tied to the adequacy of the models and to the speculation that no additional unmeasured covariates that might explain the treatment effects differ between the groups.

11.8 EXERCISES

Conceptual Exercises

1. Alcohol Metabolism. The subjects in the study were given alcohol on two consecutive days. On one of the days it was administered orally and on the other it was administered intravenously. The type of administration given on the first day was decided by a random mechanism. Why was this precaution taken?

2. Alcohol Metabolism. Here are two models for explaining the mean first-pass metabolism:

$$\text{Model 1: } \beta_0 + \beta_1 \text{gast} + \beta_2 \text{fem}$$

$$\text{Model 2: } \beta_0 + \beta_1 \text{gast} + \beta_2 \text{gast} \times \text{fem}.$$

(a) Why are there no formal tools for comparing these two models? (b) For a given value of gastric activity, what is the mean first-pass metabolism for men minus the mean first-pass metabolism for women (i) from Model 1? (ii) from Model 2?

3. Alcohol Metabolism. What would be the meaning of a third-order interactive effect of gastric activity, sex, and alcoholism on the mean first-pass metabolism?

4. Blood–Brain Barrier. (a) How should rats have been randomly assigned to treatment groups? How many treatment groups were there? What is the name of this type of experimental design and this type of treatment structure? (b) How should rats have been randomly assigned to treatments if the researchers suspected that the sex of the rat might be associated with the response? What is the name of this type of experimental design?

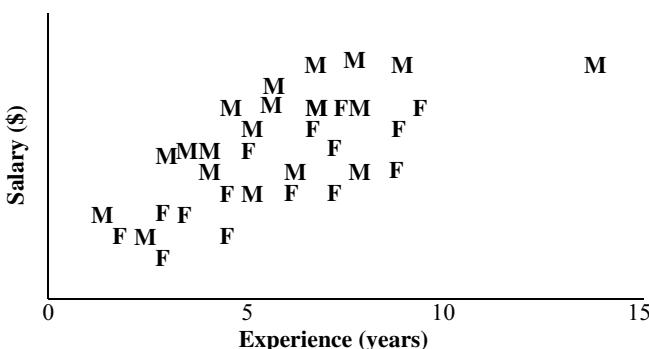
5. Blood–Brain Barrier. The residual plot in Display 11.6 contains two distinct groups of points: a cluster on the left and a cluster on the right. (a) Why is this? (b) Does it imply any problems with the model?

6. Robustness and resistance are different properties. (a) What is the difference? (b) Why are they both relevant when the response distribution is “long-tailed”?

7. Display 11.19 shows a hypothetical scatterplot of salary versus experience, with different codes for males and females. The male and female slopes differ significantly if the male with the most experience is included, but not if he is excluded. What course of action should be taken, and why?

DISPLAY 11.19

Hypothetical scatterplot of salary versus experience for males and females



8. (a) Why does a case with large leverage have the *potential* to be influential? Why is it not necessarily influential? (b) Draw a hypothetical scatterplot of Y versus a single X , where one observation

has a high leverage but is not influential. (c) Draw a hypothetical scatterplot of Y versus a single X , where one observation has a high leverage and is influential.

9. Suppose it is desired to obtain partial residuals in order to plot Y versus X_2 after getting the effect of X_1 out of the way. The first task is to fit the regression of Y on X_1 and X_2 . Let the estimated coefficients be $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$, and let the residuals from this fit be represented by res_i . The definition of the i th partial residual is $pres_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i}$. The alternative computational formula is $res_i + \hat{\beta}_2 X_{2i}$. Why are these formulas equivalent?

Computational Exercises

10. **Pollen Removal.** Reconsider the pollen removal data in Exercise 3.27. (a) Draw a coded scatterplot of the log of the proportion of pollen removed relative to the proportion unremoved (if p is the proportion removed, take $Y = \log[p/(1 - p)]$) versus the log of the duration of visit, with a code to distinguish queens from workers. (b) Fit the regression of Y on the two explanatory variables and their interaction, and obtain a residual plot. Does the residual plot indicate any problems? (c) Obtain a set of case influence statistics. Are there any problem observations? What is the most advisable course of action? (d) Does a significant interaction appear to exist, or can the simpler parallel regression lines model be used?

11. **Chernobyl Fallout.** The data in Display 11.20 are the cesium ($^{134}\text{Cs} + ^{137}\text{Cs}$) concentrations (in Bq/kg) in soil and in mushrooms at 17 wooded locations in Umbria, Central Italy, from August 1986 to November 1989. Researchers wished to investigate the cesium transfer from contaminated soil to plants—after the Chernobyl nuclear power plant accident in April 1986—by describing the distribution of the mushroom concentration as a function of soil concentration. (a) Obtain a set of case influence statistics from the simple linear regression fit of mushroom concentration on soil concentration. What do these indicate about case number 17? (b) Repeat part (a) after taking the logarithms of both variables. (Data from R. Borio et al., “Uptake of Radiocesium by the Mushrooms,” *Science of the Total Environment* 106 (1991): 183–90.)

DISPLAY 11.20

Cesium ($^{134}\text{Cs} + ^{137}\text{Cs}$) concentrations (in Bq/kg) in soil and in mushrooms at 17 locations in Italy, after the Chernobyl nuclear power plant accident; first 5 of 17 rows

Mushroom	Soil
1	33
9	55
14	138
17	319
20	415

12. **Brain Weights.** Reconsider the brain weight data of Display 9.4. (a) Fit the regression of brain weight on body weight, gestation, and log litter size, using no transformations. Obtain a set of case-influence statistics. Is any mammal influential in this fit? (b) Refit the regression without the influential observation, and obtain the new set of case influence statistics. Are there any influential observations from this fit? (c) What lessons about the connection between the need for a log transformation and influence can be discerned?

13. **Brain Weights.** Identifying which mammals have larger brain weights than were predicted by the regression model might point the way to further variables that can be examined. Fit the regression of log brain weight on log body weight, log gestation, and log litter size, and compute the studentized

residuals. Which mammals have substantially larger brain weights than were predicted by the model? Do any mammals have substantially smaller brain weights than were predicted by the model?

14. Corn Yield and Rainfall. Reconsider the data in Exercise 9.15. Fit the regression of corn yield on rainfall, rainfall-squared, and year. (a) Obtain the partial residuals of corn yield, adjusted for rainfall, and plot them versus year. (b) Obtain the augmented partial residual of corn yield, adjusted for year, and plot these values versus rainfall. (c) In your opinion, do these plots provide any clarification over the ordinary scatterplots?

15. Election Fraud. Reconsider the disputed election data from Exercise 8.20. (a) Draw a scatterplot of the Democratic percentage of absentee votes versus the Democratic percentage of machine votes for the 22 elections. Fit the simple linear regression of Democratic percentage of absentee votes on the Democratic percentage of machine votes and include the line on the plot. (b) Find the *internally* studentized residual for case number 22 from this fit. (c) Find the *externally* studentized residual for case number 22 from this fit. (d) Are the externally and internally studentized residuals for case 22 very different? If so, what can explain the difference? (e) What do the studentized residuals for case 22 indicate about the unusualness of the absentee ballot percentage for election 22 relative to the pattern of absentee and machine percentages established from the other 21 elections?

16. First-Pass Metabolism. Calculate the leverage, the studentized residual, and Cook's Distance for the 32nd case. Use the model with gastric activity, a sex indicator variable, and the interaction of these two.

17. Blood-Brain Barrier. (a) Using the data in Display 11.4, compute "jittered" versions of treatment, days after inoculation, and an indicator variable for females by adding small random numbers to each (uniform random numbers between -0.15 and 0.15 work well). (b) Obtain a matrix of the correlation coefficients among the same five variables (not jittered!). (c) In pencil, write the relevant correlation (two digits is enough) in a corner of each of the scatterplots in the matrix of scatterplots. (d) On the basis of this, what can be said about the relationship between the covariates (sex and days after inoculation), the response, and the design variables (treatment and sacrifice time)?

18. Blood-Brain Barrier. Using the data in Display 11.4, fit the regression of the log response (brain tumor-to-liver antibody ratio) on all covariates, the treatment indicator, and sacrifice time, treated as a factor with four levels (include three indicator variables, for sacrifice time = 3, 24, and 72 hours). (a) Obtain a set of case influence statistics, including a measure of influence, the leverage, and the studentized residual. (b) Discuss whether any influential observations or outliers occur with respect to this fit.

19. Blood-Brain Barrier. (a) Using the data in Display 11.4, fit the regression of the log response (brain tumor-to-liver antibody ratio) on an indicator variable for treatment and on sacrifice time treated as a factor with four levels (include three indicator variables, for sacrifice time = 3, 24, and 72 hours). Use the model to find the estimated mean of the log response at each of the eight treatment combinations (all combinations of the two infusions and the four sacrifice times). (b) Let X represent log of sacrifice time. Fit the regression of the log response on an indicator variable for treatment, X , X^2 , and X^3 . Use the estimated model to find the estimated mean of the log response at each of the eight treatment combinations. (c) Why are the answers to parts (a) and (b) the same?

20. Warm-Blooded T. Rex? The data in Display 11.21 are the isotopic composition of structural bone carbonate (X) and the isotopic composition of the coexisting calcite cements (Y) in 18 bone samples from a specimen of the dinosaur *Tyrannosaurus rex*. Evidence that the mean of Y is positively associated with X was used in an argument that the metabolic rate of this dinosaur resembled warm-blooded more than cold-blooded animals. (Data from R. E. Barrick and W. J. Showers, "Thermophysiology of *Tyrannosaurus rex*: Evidence from Oxygen Isotopes," *Science* 265 (1994): 222–24.) (a) Examine the effects on the p -value for significance of regression and on

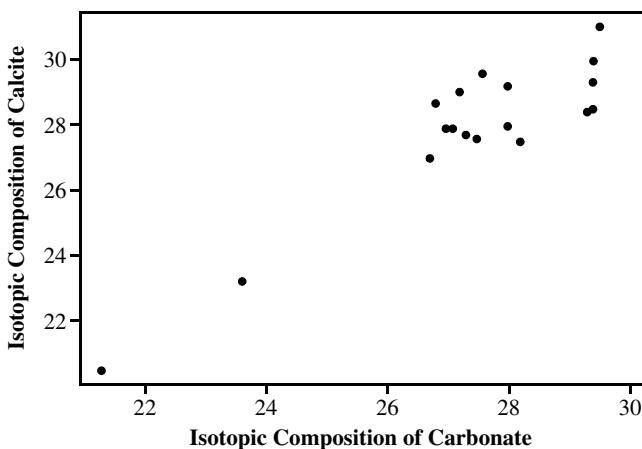
R -squared of deleting (i) the case with the smallest value of X , and (ii) the two cases with the smallest values of X . (b) Why does R -squared change so much? (c) Compute the case influence statistics, and discuss interesting cases. (d) Recompute the case statistics when the case with the smallest X is deleted. (e) Comment on the differences in the two sets of case statistics. Why may pairs of influential observations not be found with the usual case influence statistics? (f) What might one conclude about the influence of the two unusual observations in this data set?

DISPLAY 11.21

Isotopic composition of carbonate and of calcite cements in 18 samples of bone from a *Tyrannosaurus rex* specimen

Carbonate Calcite

21.3	20.5
23.6	23.2
26.7	27.0
26.8	28.7
27.0	27.9
27.1	27.9
27.2	29.0
27.3	27.7
27.5	27.6
27.6	29.6
28.0	28.0
28.0	29.2
28.2	27.5
29.3	28.4
29.4	28.5
29.4	29.3
29.4	30.0
29.5	31.0



21. Calculus Problem. The weighted least squares problem in multiple linear regression is to find the parameter values that minimize the weighted sum of squares,

$$\text{SS}_w(\beta_0, \beta_1, \dots, \beta_p) = \sum_{i=1}^n w_i (Y_i - \beta_0 - \beta_1 X_{1i} - \dots - \beta_p X_{pi})^2,$$

with all $w_i > 0$. (a) Setting the partial derivatives of SS_w with respect to each of the parameters equal to zero, show that the solutions must satisfy this set of normal equations:

$$\begin{aligned} \beta_0 \sum w_i + \beta_1 \sum w_i X_{1i} + \beta_2 \sum w_i X_{2i} + \dots + \beta_p \sum w_i X_{pi} &= \sum w_i Y_i \\ \beta_0 \sum w_i X_{1i} + \beta_1 \sum w_i X_{1i}^2 + \beta_2 \sum w_i X_{1i} X_{2i} + \dots + \beta_p \sum w_i X_{1i} X_{pi} &= \sum w_i X_{1i} Y_i \\ \beta_0 \sum w_i X_{2i} + \beta_1 \sum w_i X_{2i} X_{1i} + \beta_2 \sum w_i X_{2i}^2 + \dots + \beta_p \sum w_i X_{2i} X_{pi} &= \sum w_i X_{2i} Y_i \\ \vdots &\quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ \beta_0 \sum w_i X_{pi} + \beta_1 \sum w_i X_{pi} X_{1i} + \beta_2 \sum w_i X_{pi} X_{2i} + \dots + \beta_p \sum w_i X_{pi}^2 &= \sum w_i X_{pi} Y_i. \end{aligned}$$

(b) Show that solutions to the normal equations minimize SS.

Data Problems

22. Deforestation and Debt. It has been theorized that developing countries cut down their forests to pay off foreign debt. Two researchers examined this belief using data from 11 Latin American nations. (Data from R. T. Gullison and E. C. Losos, “The Role of Foreign Debt in Deforestation in Latin America,” *Conservation Biology* 7(1) (1993): 140–7.) The data on debt, deforestation, and population appear in Display 11.22. Does the evidence significantly support the theory that debt causes deforestation? Does debt exert any effect after the effect of population on deforestation is accounted for? Describe the effect of debt, after accounting for population.

DISPLAY 11.22 Foreign debt, annual deforestation area, and population for 11 Latin American countries

Country	Debt (millions of dollars)	Deforestation (thousands of hectares)	Population (thousands of people)
Brazil	86,396	12,150	128,425.0
Mexico	79,613	2,680	74,194.5
Ecuador	6,990	1,557	8,750.5
Colombia	10,101	1,500	27,254.0
Venezuela	24,870	1,430	16,170.5
Peru	10,707	1,250	18,496.5
Nicaragua	3,985	550	3,021.5
Argentina	36,664	400	29,400.5
Bolivia	3,810	300	5,970.5
Paraguay	1,479	250	3,424.5
Costa Rica	3,413	90	2,439.5

23. Air Pollution and Mortality. Does pollution kill people? Data in one early study designed to explore this issue came from five Standard Metropolitan Statistical Areas (SMSA) in the United States, obtained for the years 1959–1961. (Data from G. C. McDonald and J. A. Ayers, “Some Applications of the ‘Chernoff Faces’: A Technique for Graphically Representing Multivariate Data,” in *Graphical Representation of Multivariate Data*, New York: Academic Press, 1978.) Total age-adjusted mortality from all causes, in deaths per 100,000 population, is the response variable. The explanatory variables listed in Display 11.23 include mean annual precipitation (in inches); median number of school years completed, for persons of age 25 years or older; percentage of 1960 population that is nonwhite; relative pollution potential of oxides of nitrogen, NO_X; and relative pollution potential of sulfur dioxide, SO₂. “Relative pollution potential” is the product of the tons emitted per day per square kilometer and a factor correcting for SMSA dimension and exposure. The first three explanatory variables are a subset of climate and socioeconomic variables in the original data set. (Note: Two cities—Lancaster and York—are heavily populated by members of the Amish religion,

DISPLAY 11.23 Air pollution and mortality data for 5 U.S. cities, 1959–1961; first 5 of 60 rows

City	Mortality	Precipitation	Education	Nonwhite	NO _x	SO ₂
San Jose, CA	790.73	13	12.2	3.0	32	3
Wichita, KS	823.76	28	12.1	7.5	2	1
San Diego, CA	839.71	10	12.1	5.9	66	20
Lancaster, PA	844.05	43	9.5	2.9	7	32
Minneapolis, MN	857.62	25	12.1	2.0	11	26

who prefer to teach their children at home. The lower years of education for these two cities do not indicate a social climate similar to other cities with similar years of education.) Is there evidence that mortality is associated with either of the pollution variables, after the effects of the climate and socioeconomic variables are accounted for? Analyze the data and write a report of the findings, including any important limitations of this study. (*Hint:* Consider looking at case-influence statistics.)

24. Natal Dispersal Distances of Mammals. Natal dispersal distances are the distances that juvenile animals travel from their birthplace to their adult home. An assessment of the factors affecting dispersal distances is important for understanding population spread, recolonization, and gene flow—which are central issues for conservation of many vertebrate species. For example, an understanding of dispersal distances will help to identify which species in a community are vulnerable to the loss of connectedness of habitat. To further the understanding of determinants of natal dispersal distances, researchers gathered data on body weight, diet type, and maximum natal dispersal distance for various animals. Shown in Display 11.24 are the first 6 of 64 rows of data on mammals. (Data from G. D. Sutherland et al., “Scaling of Natal Dispersal Distances in Terrestrial Birds and Mammals,” *Conservation Ecology* 4(1) (2000): 16.) Analyze the data to describe the distribution of maximum dispersal distance as a function of body mass and diet type. Write a summary of statistical findings.

DISPLAY 11.24 Natal dispersal distances and explanatory variables for 64 mammals; first 6 of 64 rows

Species	Body mass (kg)	Diet type	Maximum dispersal distance (km)
1. <i>Didelphis virginianus</i>	2.41	Omnivore	5.15
2. <i>Phascogale tapotafa</i>	0.17	Carnivore	6.80
3. <i>Trichosurus vulpecula</i>	2.93	Carnivore	12.80
4. <i>Sorex araneus</i>	0.004	Carnivore	0.87
5. <i>Scapanus townsendii</i>	0.15	Omnivore	0.86
6. <i>Ursus americanus</i>	104.45	Omnivore	225.00

25. Ingestion Rates of Deposit Feeders. The ingestion rates and organic consumption percentages of deposit feeders were considered in Exercise 9.21. The data set ex1125 repeats these data, but this time with three additional bivalve species included (the last three). The researcher wished to see if ingestion rate is associated with the percentage of organic matter in food, after accounting for animal weight, but was unsure about whether bivalves should be included in the analysis. Analyze the data to address this question of interest. (Data from L. M. Cammen, “Ingestion Rate: An Empirical Model for Aquatic Deposit Feeders and Detritivores,” *Oecologia* 44 (1980): 303–10.)

26. Metabolism and Lifespan of Mammals. Use the data from Exercise 8.26 to describe the distribution of mammal lifespan as a function of metabolism rate, after accounting for the effect of body mass. One theory holds that for a given body size, animals that expend less energy per day (lower metabolic rate) will tend to live longer.

Answers to Conceptual Exercises

1. By randomly determining order, the researchers avoid bias in determining first-pass metabolism that would occur if an order effect existed.
2. (a) Neither of the models is a subset of the others, so it is impossible to test a term in a “full model” to see whether the “reduced model” does just as well. (Incidentally, the models explain the data about equally well.) (b) (i) β_2 (ii) β_2gast .
3. It would mean that the effect of gastric AD activity on first-pass metabolism differed between males and females, and that the amount of the sex difference differed between alcoholics and

nonalcoholics. (Two-factor interactions are hard enough to describe in words. Three- and higher-factor interactions typically involve very long and confusing sentences. A theory without interactions is obviously simpler than one with interactions.)

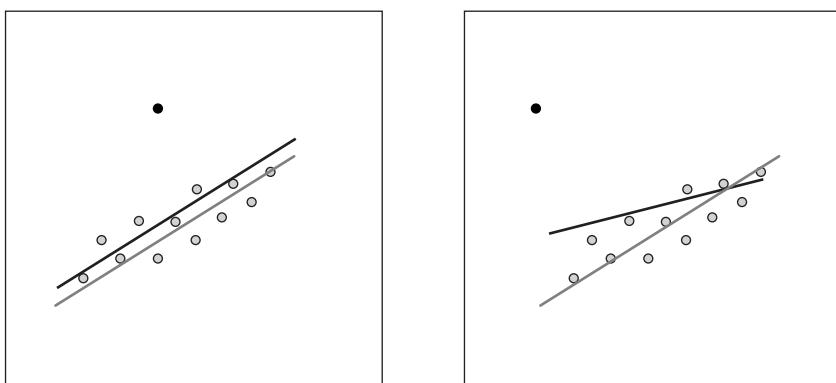
4. (a) Rats should have been randomly assigned to one of eight groups, corresponding to the eight combinations of treatment and sacrifice time. This is a completely randomized design with factorial (2×4) treatment structure. (b) If the response is suspected to be related to the sex of the rat, a randomized block experiment should be performed. The procedure in part (a) can be followed separately for male and female rats.

5. (a) The fitted values are noticeably larger for the rats in the groups with longer sacrifice times. (b) No.

6. (a) *Robustness* describes the extent to which inferential statements are correct when specific assumptions are violated. *Resistance* describes the extent to which the results remain unchanged when a small portion of the data is changed, perhaps drastically (and therefore describes the extent to which individual observations can be influential). (b) When the distribution is long-tailed, the robustness against departures from normality cannot be guaranteed. Put another way, there are likely to be outliers, which can have undue influence on the results, since the least squares method is not resistant.

7. In the absence of further knowledge, it is safest to exclude the very experienced male from the data set and restrict conclusions about the differences between male and female salaries to individuals with 10 years of experience or fewer. It may be that males and females have different slopes over the wider range of experience, or it may be that the straight line is not an adequate model over the wider range of experiences. There is certainly insufficient data in the more-than-10-years range to resolve this issue.

8. (a) A large leverage indicates that a case occupies a position in the “ X -space” that is not densely populated. It therefore plays a large role in shaping the estimated regression model in that region. Since it does not share its role (much) with other nearby points, it must draw the regression surface close to it. For this reason it has a high potential for influence. If, however, the fit of the model without that point is about the same as the fit of the model with it, it is not influential.



(b) The case with high leverage (\bullet) exerts no influence on the slope. The line without the case ($-$) only has its intercept changed when the case is included (--).

(c) When the case with high leverage has its explanatory variable value outside those of the other cases, the slope of the regression can be changed dramatically.

9. Since $res_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}$, the alternative calculating formula is $Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} + \hat{\beta}_2 X_{2i}$. The last two terms cancel, leaving the original definition.

Strategies for Variable Selection

There are two good reasons for paring down a large number of explanatory variables to a smaller set. The first reason is somewhat philosophical: Simplicity is preferable to complexity. Thus, redundant and unnecessary explanatory variables should be excluded on principle. The second reason is more concrete: Unnecessary terms in the model yield less precise inferences.

Various statistical tools are available for choosing a good subset from a large pool of explanatory variables. Discussed in this chapter are sequential variable-selection techniques, and comparisons among all possible subsets through examination of C_p and the Bayesian Information Criterion. The most important practical lessons are that the variable selection process should be sensitive to the objectives of the study and that the particular subset chosen is relatively unimportant.

12.1 CASE STUDIES

12.1.1 State Average SAT Scores—An Observational Study

When, in 1982, average Scholastic Achievement Test (SAT) scores were first published on a state-by-state basis in the United States, the huge variation in the scores was a source of great pride for some states and of consternation for others. Average scores ranged from a low of 790 (out of a possible 1,600) in South Carolina to a high of 1,088 in Iowa. This 298-point spread dwarfed the 20-year national decline of 80 points. Two researchers set out to “assess the extent to which the compositional/demographic and school-structural characteristics are implicated in SAT differences.” (Data from B. Powell and L. C. Steelman, “Variations in State SAT Performance: Meaningful or Misleading?” *Harvard Educational Review* 54(4) (1984): 389–412.)

Display 12.1 contains state averages of the total SAT (verbal + quantitative) scores, along with six variables that may be associated with the SAT differences among states. Some explanatory variables come from the Powell and Steelman article, while others were obtained from the College Entrance Examination Board (by Robert Powers). The variables are the following: *takers* is the percentage of the total eligible students (high school seniors) in the state who took the exam; *income* is the median income of families of test-takers, in hundreds of dollars; *years* is the average number of years that the test-takers had formal studies in social sciences, natural sciences, and humanities; *public* is the percentage of the test-takers who attended public secondary schools; *expend* is the total state expenditure on secondary schools, expressed in hundreds of dollars per student; and *rank* is the median percentile ranking of the test-takers within their secondary school classes.

Notice that the states with high average SATs had low percentages of takers. One reason is that these are mostly midwestern states that administer other tests to students bound for college in-state. Only their best students planning to attend college out of state take the SAT exams. As the percentage of takers increases for other states, so does the likelihood that the takers include lower-qualified students. After accounting for the percentage of students who took the test and the median class rank of the test-takers (to adjust, somewhat, for the selection bias in the samples from each state), which variables are associated with state SAT scores? After accounting for the percentage of takers and the median class rank of the takers, how do the states rank? Which states perform best for the amount of money they spend?

Statistical Conclusion

The percentage of eligible students taking the test and the median class rank of the students taking the test explain 81.5% of the variation in state average test scores. This confirms the expectation that much of the between-states variation in SAT scores can be explained by the relative quality of the students within the state who decide to take the test. After the percentage of students in a state who take the test and the median class rank of these students are accounted for, convincing

DISPLAY 12.1 Average SAT scores by U.S. state in 1982, and possible associated factors

Rank	State	SAT	Takers	Income	Years	Public	Expend	Rank
1	Iowa	1,088	3	326	16.79	87.8	25.60	89.7
2	South Dakota	1,075	2	264	16.07	86.2	19.95	90.6
3	North Dakota	1,068	3	317	16.57	88.3	20.62	89.8
4	Kansas	1,045	5	338	16.30	83.9	27.14	86.3
5	Nebraska	1,045	5	293	17.25	83.6	21.05	88.5
6	Montana	1,033	8	263	15.91	93.7	29.48	86.4
7	Minnesota	1,028	7	343	17.41	78.3	24.84	83.4
8	Utah	1,022	4	333	16.57	75.2	17.42	85.9
9	Wyoming	1,017	5	328	16.01	97.0	25.96	87.5
10	Wisconsin	1,011	10	304	16.85	77.3	27.69	84.2
11	Oklahoma	1,001	5	358	15.95	74.2	20.07	85.6
12	Arkansas	999	4	295	15.49	86.4	15.71	89.2
13	Tennessee	999	9	330	15.72	61.2	14.58	83.4
14	New Mexico	997	8	316	15.92	79.5	22.19	83.7
15	Idaho	995	7	285	16.18	92.1	17.80	85.9
16	Mississippi	988	3	315	16.76	67.9	15.36	90.1
17	Kentucky	985	6	330	16.61	71.4	15.69	86.4
18	Colorado	983	16	333	16.83	88.3	26.56	81.8
19	Washington	982	19	309	16.23	87.5	26.53	83.2
20	Arizona	981	11	314	15.98	80.9	19.14	84.3
21	Illinois	977	14	347	15.80	74.6	24.41	78.7
22	Louisiana	975	5	394	16.85	44.8	19.72	82.9
23	Missouri	975	10	322	16.42	67.7	20.79	80.6
24	Michigan	973	10	335	16.50	80.7	24.61	81.8
25	West Virginia	968	7	292	17.08	90.6	18.16	86.2
26	Alabama	964	6	313	16.37	69.6	13.84	83.9
27	Ohio	958	16	306	16.52	71.5	21.43	79.5
28	New Hampshire	925	56	248	16.35	78.1	20.33	73.6
29	Alaska	923	31	401	15.32	96.5	50.10	79.6
30	Nevada	917	18	288	14.73	89.1	21.79	81.1
31	Oregon	908	40	261	14.48	92.1	30.49	79.3
32	Vermont	904	54	225	16.50	84.2	20.17	75.8
33	California	899	36	293	15.52	83.0	25.94	77.5
34	Delaware	897	42	277	16.95	67.9	27.81	71.4
35	Connecticut	896	69	287	16.75	76.8	26.97	69.8
36	New York	896	59	236	16.86	80.4	33.58	70.5
37	Maine	890	46	208	16.05	85.7	20.55	74.6
38	Florida	889	39	255	15.91	80.5	22.62	74.6
39	Maryland	889	50	312	16.90	80.4	25.41	71.5
40	Virginia	888	52	295	16.08	88.8	22.23	72.4
41	Massachusetts	888	65	246	16.79	80.7	31.74	69.9
42	Pennsylvania	885	50	241	17.27	78.6	27.98	73.4
43	Rhode Island	877	59	228	16.67	79.7	25.59	71.4
44	New Jersey	869	64	269	16.37	80.6	27.91	69.8
45	Texas	868	32	303	14.95	91.7	19.55	76.4
46	Indiana	860	48	258	14.39	90.2	17.93	74.1
47	Hawaii	857	47	277	16.40	67.6	21.21	69.9
48	North Carolina	827	47	224	15.31	92.8	19.92	75.3
49	Georgia	823	51	250	15.55	86.5	16.52	74.0
50	South Carolina	790	48	214	15.42	88.1	15.60	74.0

evidence exists that both state expenditures (one-sided p -value < 0.0001) and years of formal study in social sciences, natural sciences, and humanities (one-sided p -value = 0.0005) are associated with SAT averages. Alaska had a substantially higher expenditure than other states and was excluded from the analysis.

The ranking of states after accounting for the relative quality of students in the state who take the exam (as represented by the percentage of takers and the median class rank of the students taking the exam) is shown on the left side of Display 12.2. Here, the states are ordered according to the size of their residuals in the regression of SAT scores on percentage of takers and median class rank. For example, the SAT average for New Hampshire is 49 points higher than the estimated mean SAT for states with the same percentage of students taking the exam and median class rank. This ranking and the ranking based on raw averages show some dramatic differences. South Dakota ranks second in its raw SAT scores, for example (see Display 12.1), but only 2% of its eligible students took the exam. After percentage of takers and median class rank are accounted for, South Dakota ranks 24th.

The ranking of states after additionally accounting for state expenditure is shown on the right side of Display 12.2. This ranking provides a means of gauging how states perform for the amount of money they spend (but see the cautionary notes in the next subsection regarding scope of inference). For example, Oregon ranks 31st in its unadjusted average SAT. But after accounting for the percentage of takers, their median class rankings, and the state's expenditure per student, it ranks 46th, indicating that its students performed below what might have been expected given the amount of money the state spends on education.

Scope of Inference

The participating students were *self-selected*; that is, each student included in a state's sample decided to take the SAT, as opposed to being a randomly selected student. Therefore, a state's average does not represent all its eligible students. Including the percentage of takers and the median class rank of takers in the regression models adjusts for this difference, but only to the extent that these variables adequately account for the relative quality of students taking the test. No check on this assumption is possible with the existing data. In addition, these are observational data, for which many possible confounding factors exist, so the statistical statements of significance (as for the effect of expenditure, for example) do not imply causation.

12.1.2 Sex Discrimination in Employment—An Observational Study

Display 12.3 lists data on employees from one job category (skilled, entry-level clerical) of a bank that was sued for sex discrimination. These are the same 32 male and 61 female employees, hired between 1965 and 1975, who were considered in Section 1.1.2. The measurements are of annual salary at time of hire, salary as of March 1977, sex (1 for females and 0 for males), seniority (months since first hired), age (months), education (years), and work experience prior to employment with the bank (months).

DISPLAY 12.2

State SAT scores after adjustment (in points above or below average)

SATs adjusted for % taking exam and their median class rank			SATs adjusted for % taking exam, rank, and expenditure		
Rank	State	Adjusted SAT average	Rank	State	Adjusted SAT average
1	New Hampshire	49	1	New Hampshire	67
2	Iowa	41	2	Tennessee	49
3	Montana	38	3	Vermont	39
4	Connecticut	38	4	Connecticut	28
5	Washington	34	5	Nebraska	22
6	Minnesota	34	6	Virginia	20
7	Wisconsin	31	7	Maine	18
8	Colorado	31	8	Arizona	18
9	New York	29	9	Minnesota	17
10	Kansas	29	10	North Dakota	17
11	Massachusetts	27	11	Illinois	16
12	Illinois	25	12	Ohio	15
13	Nebraska	24	13	Washington	15
14	Vermont	21	14	Iowa	14
15	North Dakota	20	15	Colorado	11
16	Tennessee	16	16	Idaho	10
17	Delaware	13	17	Utah	10
18	Maryland	12	18	Missouri	8
19	Virginia	11	19	Maryland	7
20	Ohio	11	20	New Mexico	5
21	New Mexico	8	21	South Dakota	4
22	Rhode Island	8	22	Indiana	4
23	New Jersey	8	23	Wisconsin	3
24	South Dakota	7	24	Rhode Island	3
25	Arizona	6	25	Kentucky	1
26	Pennsylvania	4	26	Montana	0
27	Missouri	4	27	Florida	-1
28	Oregon	3	28	Kansas	-1
29	Maine	2	29	Hawaii	-5
30	Michigan	-1	30	Delaware	-5
31	Wyoming	-2	31	Alabama	-5
32	Utah	-3	32	Massachusetts	-6
33	Idaho	-5	33	New Jersey	-8
34	Florida	-6	34	Oklahoma	-11
35	California	-6	35	New York	-13
36	Oklahoma	-14	36	Arkansas	-13
37	Hawaii	-19	37	Pennsylvania	-14
38	Kentucky	-23	38	Michigan	-14
39	Indiana	-24	39	California	-18
40	Nevada	-28	40	West Virginia	-19
41	West Virginia	-32	41	Texas	-22
42	Louisiana	-33	42	Georgia	-23
43	Arkansas	-34	43	Nevada	-25
44	Alabama	-38	44	Wyoming	-27
45	Texas	-40	45	Louisiana	-28
46	Georgia	-58	46	Oregon	-29
47	Mississippi	-60	47	Mississippi	-40
48	North Carolina	-61	48	North Carolina	-42
49	South Carolina	-94	49	South Carolina	-55

DISPLAY 12.3

Sex discrimination data

Beginning salary	1977 salary	Fsex (1 = F)	Seniority	Age	Education	Experience
5,040	12,420	0	96	329	15	14
6,300	12,060	0	82	357	15	72
6,000	15,120	0	67	315	15	35.5
6,000	16,320	0	97	354	12	24
6,000	12,300	0	66	351	12	56
6,840	10,380	0	92	374	15	41.5
8,100	13,980	0	66	369	16	54.5
6,000	10,140	0	82	363	12	32
6,000	12,360	0	88	555	12	252
6,900	10,920	0	75	416	15	132
6,900	10,920	0	89	481	12	175
5,400	12,660	0	91	331	15	17.5
6,000	12,960	0	66	355	15	64
6,000	12,360	0	86	348	15	25
5,100	8,940	1	95	640	15	165
4,800	8,580	1	98	774	12	381
5,280	8,760	1	98	557	8	190
5,280	8,040	1	88	745	8	90
4,800	9,000	1	77	505	12	63
4,800	8,820	1	76	482	12	6
5,400	13,320	1	86	329	15	24
5,520	9,600	1	82	558	12	97
5,400	8,940	1	88	338	12	26
5,700	9,000	1	76	667	12	90
3,900	8,760	1	98	327	12	0
4,800	9,780	1	75	619	12	144
6,120	9,360	1	78	624	12	208.5
5,220	7,860	1	70	671	8	102
5,100	9,660	1	66	554	8	96
4,380	9,600	1	92	305	8	6.25
4,290	9,180	1	69	280	12	5
5,400	9,540	1	66	534	15	122
4,380	10,380	1	92	305	12	0
5,400	8,640	1	65	603	8	173
5,400	11,880	1	66	302	12	26
4,500	12,540	1	96	366	8	52
5,400	8,400	1	70	628	12	82
5,520	8,880	1	67	694	12	196
5,640	10,080	1	90	368	12	55
4,800	9,240	1	73	590	12	228
5,400	8,640	1	66	771	8	228
4,500	7,980	1	80	298	12	8
5,400	11,940	1	77	325	12	38
5,400	9,420	1	72	589	15	49
6,300	9,780	1	66	394	12	86.5
5,160	10,680	1	87	320	12	18
5,100	11,160	1	98	571	15	115
4,800	8,340	1	79	602	8	70

DISPLAY 12.3

Sex discrimination data—continued

Beginning salary	1977 salary	Fsex (1 = F)	Seniority	Age	Education	Experience
5,400	9,600	1	98	568	12	244
4,020	9,840	1	92	528	10	44
4,980	8,700	1	74	718	8	318
5,280	9,780	1	88	653	12	107
5,700	8,280	1	65	714	15	241
4,800	8,340	1	87	647	12	163
4,800	13,560	1	82	338	12	11
5,700	10,260	1	82	362	15	51
4,380	9,720	1	93	303	12	4.5
4,380	10,500	1	89	310	12	0
5,400	10,680	0	88	359	12	38
5,400	11,640	0	96	474	12	113
5,100	7,860	0	84	535	12	180
6,600	11,220	0	66	369	15	84
5,100	8,700	0	97	637	12	315
6,600	12,240	0	83	536	15	215.5
5,700	11,220	0	94	392	15	36
6,000	12,180	0	91	364	12	49
6,000	11,580	0	83	521	15	108
6,000	8,940	0	80	686	12	272
6,000	10,680	0	87	364	15	56
4,620	11,100	0	77	293	12	11.5
5,220	10,080	0	85	344	12	29
6,600	15,360	0	83	340	15	64
5,400	12,600	0	78	305	12	7
6,000	8,940	0	78	659	8	320
5,400	9,480	0	88	690	15	359
6,000	14,400	0	96	402	16	45.5
5,700	10,620	1	88	410	15	61
5,400	10,320	1	78	584	15	51
4,440	9,600	1	97	341	15	75
6,300	10,860	1	84	662	15	231
6,000	9,720	1	69	488	12	121
5,100	9,600	1	85	406	12	59
4,800	11,100	1	87	349	12	11
5,100	10,020	1	87	508	16	123
5,700	9,780	1	74	542	12	116.5
5,400	10,440	1	72	604	12	169
5,100	10,560	1	84	458	12	36
4,800	9,240	1	84	571	16	214
6,000	11,940	1	86	486	15	78.5
4,380	10,020	1	93	313	8	7.5
5,580	7,860	1	69	600	12	132.5
4,620	9,420	1	96	385	12	52
5,220	8,340	1	70	468	12	127

Did the females receive lower starting salaries than similarly qualified and similarly experienced males? After accounting for measures of performance, did females receive smaller pay increases than males?

Statistical Conclusion

The data provide convincing evidence that the median starting salary for females was lower than the median starting salary for males, even after the effects of age, education, previous experience, and time at which the job began are taken into account (one-sided p -value <0.0001). The median beginning salary for females was estimated to be only 89% of the median salary for males, after accounting for the variables mentioned above (a 95% confidence interval for the ratio of adjusted medians is 85% to 93%). There is little evidence that the pay increases for females differed from those for males (one-sided p -value = 0.27 from a t -test for the difference in mean log of average annual raise; one-sided p -value = 0.72 after further adjustment for other variables except beginning salary; one-sided p -value = 0.033 after further adjustment for other variables including beginning salary).

Inferences

Since these are observational data, the usual dictum applies: No cause-and-effect inference of discrimination can be drawn. Of course that argument disallows nearly all claims of discrimination, because supporting data are always observational. To evaluate such evidence, U.S. courts have adopted a method of burden-shifting. Plaintiffs present statistical analyses as *prima facie* evidence of discrimination. If the evidence is substantial, the burden shifts to the defendants to show either that these analyses are flawed or that alternative, nondiscriminatory factors explain the differences. If the defendants' arguments are accepted, the burden shifts back to the plaintiffs to show that the defendants' analyses are wrong or that their explanations were pretextual. If plaintiffs succeed in this, the inference of discrimination stands.

12.2 SPECIFIC ISSUES RELATING TO MANY EXPLANATORY VARIABLES

12.2.1 Objectives

Several convenient tools are available for paring down large sets of explanatory variables. But without understanding what they offer, one may incorrectly suppose that the computer is uncovering some law of nature. Like any other statistical tool, these paring tools are most helpful when used to address well-defined questions of interest. Although it may be tempting to think that finding “the important set” of explanatory variables is a question of interest, the actual set chosen is usually one of several (or many) equally good sets.

Objective I: Adjusting for a Large Set of Explanatory Variables

In the sex discrimination example the objective is to examine the effect of sex after accounting for other legitimate determinants of salary. A game plan for the analysis

is to begin by finding a subset of the other explanatory variables with a variable-selection technique, and then to add the indicator variable for sex into the model. The variable-selection techniques are entirely appropriate for this purpose. The set of explanatory variables chosen is evidently one of several (or perhaps many) equally useful sets. Although a particular explanatory variable may not be included in the final model, it has still been adjusted for, since it was given a chance to be in the model (prior to inclusion of the sex indicator). No interpretation is made of the particular set chosen, nor is any interpretation made of the coefficients. The coefficient of the sex indicator variable is interpreted, but this is straightforward: It represents the association between sex and salary *after* accounting for the effects of the other explanatory variables.

Objective II: Fishing for Explanation

In many studies, unfortunately, no such well-defined question has been posed. One question likely to be asked of the SAT data is vague: “Which variables are important in explaining state average SAT scores, after accounting for percentage of takers and median class rank?” Variable-selection techniques can be used to identify a set of explanatory variables, but the analyst may be tempted to attach meaning to the particular variables selected and to interpret coefficients—temptations not encountered in Objective I.

There are several reasons to exercise great caution when using regression for this purpose:

1. The explanatory variables chosen are not necessarily special. Inclusion or exclusion of individual explanatory variables is affected strongly by the correlations between them.
2. Interpreting the coefficients in a multiple regression with correlated explanatory variables is extremely difficult. For example, an estimated coefficient may have the opposite sign of the one suggested by a scatterplot, since a coefficient shows the effect of an explanatory variable on a response after accounting for the other variables in the model. It is often difficult to recognize what “after accounting for the other variables in the model” means when many variables are included.
3. A regression coefficient continues to be interpreted (technically) as a measure of the effect of one variable while holding all other variables fixed. Unfortunately, increases in one variable are almost always accompanied by changes in many others, so the situation described by the interpretation lies outside the experience provided by the data.
4. Finally, of course, causal interpretations from observational studies are always suspect. The selected variables may be related to confounding variables that are more directly responsible for the response.

Objective III: Prediction

If the purpose of the model is prediction, no interpretation of the particular set of explanatory variables chosen or their coefficients is needed, so there is little room for abuse. The variable-selection techniques are useful for showing a few models

that work well. From these, the researcher selects one with explanatory variables that can be obtained conveniently for future prediction.

Different Questions Require Different Attitudes Toward the Explanatory Variables

With regard to the SAT score example, a business firm looking for a place to build a new facility may have as its objective in analyzing the data a ranking of states in terms of their ability to train students. The only relevant question to this firm is whether the raw SAT averages accurately reflect the educational training. The problem of selection bias is critical, so inclusion of the percentage of SAT takers or of the median rank is essential. All other variables reflect factors that will not affect the firm's decision, so they are not relevant. This firm could fit a model using the percentage of takers as an explanatory variable and use the residuals as a way of ranking the states.

A legislative watchdog committee, on the other hand, may have the objective of determining the impact of state expenditures on state SAT scores. It might include all variables that could affect SAT scores before including expenditures. Under this arrangement, the committee would claim an effect for expenditures only when the effect could not be attributed to some other related factor.

If prediction is the purpose, all explanatory variables that can be used in making the prediction should be considered eligible for inclusion. (Prediction is not reasonable in the SAT score example, because one would not know the explanatory variables until the SAT scores themselves were available.)

12.2.2 Loss of Precision

The precision in estimating an important regression coefficient or in predicting a future response may be decreased if too many explanatory variables are included in the model. The standard error of the estimator for a multiple regression coefficient of an explanatory variable, X , can be expressed as the standard error in the simple linear regression multiplied by the square root of the product of two factors, a *mean square ratio* and a *variance inflation factor (VIF)*. The *mean square ratio* is the ratio of estimate of residual variance from the multiple and the simple regression models. If useful predictors are added in the multiple regression model then this ratio should be less than one.

Often more important, however, is the variance inflation factor,

$$VIF = 1/(1 - R_X^2)$$

where R_X^2 is the proportion of the variation in X that is explained by its linear relationship to other explanatory variables in the model. When X can be explained well by the additional variables—a condition called *multicollinearity*—the inflation factor can be very high.

Adding additional explanatory terms, therefore, may either decrease or increase precision depending on the relative effects on the *mean square ratio* and the *VIF*. These conflicting effects highlight a fundamental tension in statistical modeling. If the purpose is to estimate the coefficient of X (or to predict a future Y) as precisely

as possible, additional explanatory terms may help if those terms reduce the unexplained variation in the response, but may hurt if this reduction is insufficient to overcome the increase in the variance inflation factor. Sections 12.3 and 12.4 return to this issue, introducing tools for finding sets of explanatory variables that include “enough” but “not too many” terms.

12.2.3 A Strategy for Dealing with Many Explanatory Variables

A strategy for dealing with many explanatory variables should include the following elements:

1. Identify the key objectives.
2. Screen the available variables, deciding on a list that is sensitive to the objectives and excludes obvious redundancies.
3. Perform exploratory analysis, examining graphical displays and correlation coefficients.
4. Perform transformations as necessary.
5. Examine a residual plot after fitting a rich model, performing further transformations and considering outliers.
6. Use a computer-assisted technique for finding a suitable subset of explanatory variables, exerting enough control over the process to be sensitive to the questions of interest.
7. Proceed with the analysis, using the selected explanatory variables.

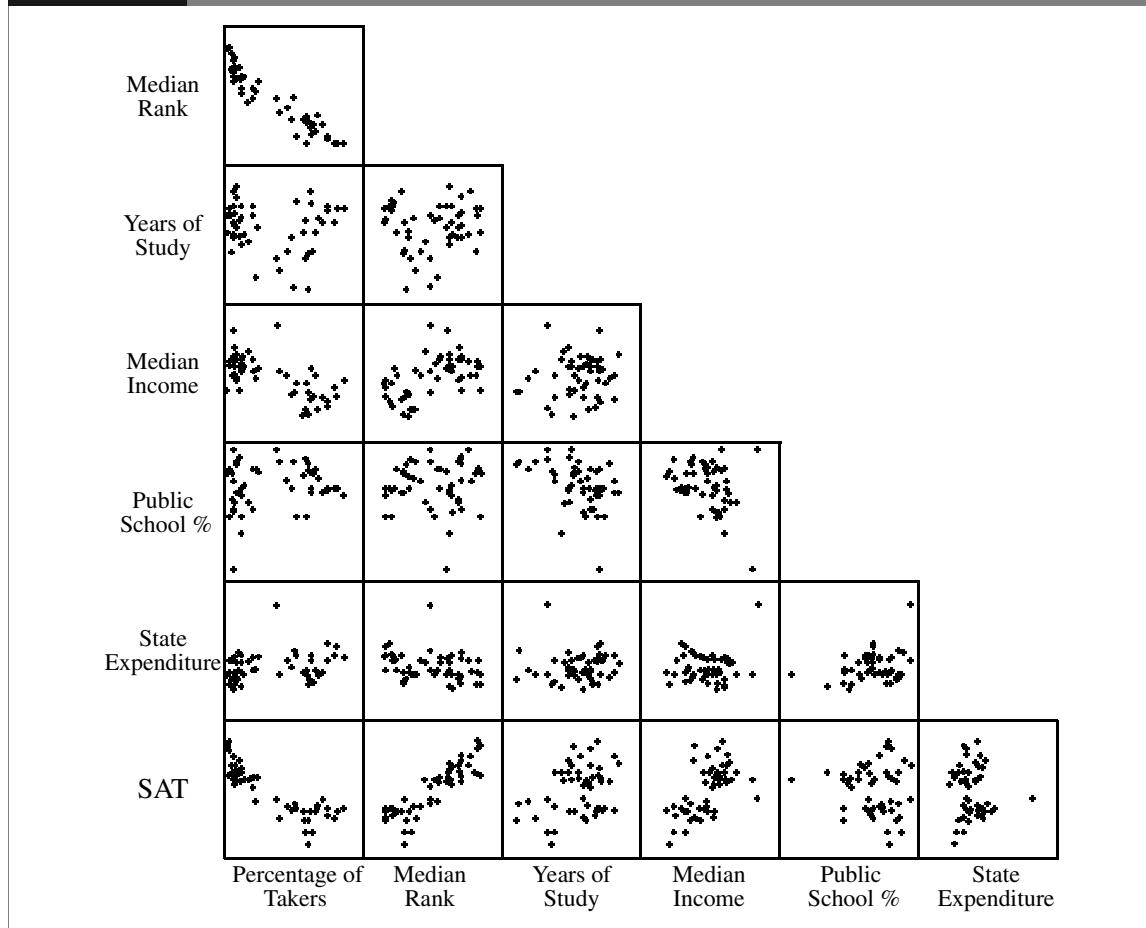
Example—Preliminary Analysis of the State SAT Data

The objectives of the SAT study were set out earlier. The variables in Display 12.1 were screened from the original source variables. (Several irrelevant variables were not retained.)

Display 12.4 shows the matrix of scatterplots for the SAT data. Examination of the bottom row indicates some nonlinearity in the relationship between SAT and percentage of takers and shows some potential outliers with respect to the variables public school percentage and state expenditure. Using the logarithm of the percentage of takers responds to the first problem, yielding a relationship that looks like a straight line. The outliers were identified: Louisiana is the state with the unusually low percentage of students attending public schools, and Alaska is the state with the very high state expenditure. These require further examination after a model has been fit.

The questions of interest call for adjustment with respect to percentage of SAT takers and median class rank of those takers in each state. When a preliminary regression of SAT was fit on both of these variables (a fit unaffected by Louisiana and Alaska), the variables were found to explain 81.5% of the variation between state SAT averages.

Partial residual plots at this stage help the researcher evaluate the effects of some of the other explanatory variables, after the first two are accounted for. Display 12.5 shows a partial residual plot of the relationship between SAT and expenditure, after the effects of percentage of takers and class rank have been

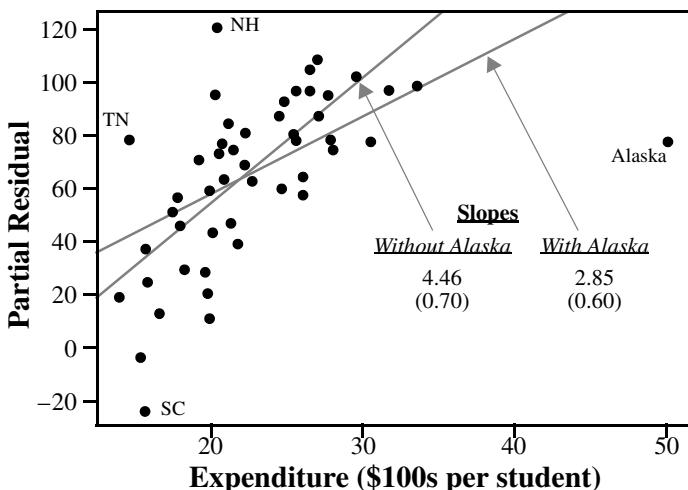
DISPLAY 12.4 Matrix of scatterplots for SAT scores and six explanatory variables

cleared out of the way. This plot has two striking features: It reveals the noticeable effect of expenditure, after the two adjustment variables are accounted for; and it demonstrates that Alaska is very influential in any fit involving expenditure. To clarify this, the diagram shows two possible straight line regression fits—one with and one without Alaska. The slopes of these two lines differ so markedly that neither value would lie within a confidence interval based on the other. Case influence statistics confirm that Alaska is influential. It has a large leverage ($h_i = 0.458$, from a statistical computer program) and does not fit well to the model that fits the other states (studentized residual = -3.28).

The prudent step is to set Alaska aside for the remainder of the analysis. It can be argued that the role of expenditure is somewhat different in that state, since the costs of heating school buildings and transporting teachers and students long distances are far greater in Alaska, and these expenses do not contribute directly to educational quality. Furthermore, too few states have similar expenditures to

DISPLAY 12.5

Partial residual plot of state average SAT scores (adjusted for percentage of students in the state who took the test and for median class rank of the students who took the test) versus state expenditure on secondary education



permit an accurate determination of the proper form of the model for expenditures greater than \$3,500 per student. A similar plot of partial residuals versus percentage of students in public schools provides no reason to be concerned about Louisiana. The effect of the percentage of students who attend public schools in a state appears to be insignificant whether Louisiana is included or not.

12.3 SEQUENTIAL VARIABLE-SELECTION TECHNIQUES

The search for a suitable subset of explanatory variables may encompass a large array of possible models. Computers facilitate the process immensely, since hundreds or even thousands of models can be analyzed in a short period of time. Yet in some instances the search may be so wide as to tie up even the best computer for an intolerable amount of time. Sequential variable selection procedures offer the option of exploring some (but not all) of the possible models.

A step in any sequential (or *stepwise*) procedure begins with a tentative *current model*. Several other models in the neighborhood of the current model (but differing from the current model in having one of its variables excluded or having one additional variable included) are examined to determine whether they are in some sense superior. If so, the best is chosen as the tentative model for the next step. Procedures may differ in their definitions of neighborhood, in their criteria for superiority, and in their initial tentative model.

12.3.1 Forward Selection

The *forward selection* procedure starts with a constant mean as its current model and adds explanatory variables one at a time until no further addition significantly improves the fit. Each forward selection step consists of two tasks:

1. Consider all models obtained by adding one more explanatory variable to the current model. For each variable not already included, calculate its “*F*-to-enter” (the extra-sum-of-squares *F*-statistic for testing its significance). Identify the variable with the largest *F*-to-enter.
2. If the largest *F*-to-enter is greater than 4 (or some other user-specified number), add that explanatory variable to form a new current model.

Tasks 1 and 2 are repeated until no additional explanatory variables can be added.

12.3.2 Backward Elimination

In the *backward elimination* method, the initial current model contains *all* possible explanatory variables. Each step in this procedure consists of two tasks:

1. For each variable in the current model, calculate the *F*-to-remove (the extra-sum-of-squares *F*-statistic for testing its significance). Identify the variable with the smallest *F*-to-remove.
2. If the smallest *F*-to-remove is 4 (or some other user-specified number) or less, then remove that explanatory variable to arrive at a new current model.

Tasks 1 and 2 are repeated until none of the remaining explanatory variables can be removed.

12.3.3 Stepwise Regression

The constant mean model with no explanatory variables is the starting current model. Each step consists of the following tasks.

1. Do one step of forward selection.
2. Do one step of backward elimination.

Tasks 1 and 2 are repeated until no explanatory variables can be added or removed.

Inclusion of Categorical Factors

Each categorical factor is represented in a regression model by a set of indicator variables. In the absence of a good reason for doing otherwise, this set should be added or removed as a single unit. Some computer packages allow this automatically.

No Universal Best Model

Forward selection, backward elimination, and stepwise regression can lead to different final models. This fact should not be alarming, since no single best model can

be expected. It does, however, emphasize a disadvantage of these methods: Only one model is presented as an answer, which may give the false impression that this model alone explains the data well. As a result, unwarrantedly narrow attention may be accorded to the particular explanatory variables in that single model.

12.3.4 Sequential Variable-Selection with the SAT Data

Display 12.6 is a skeleton of all possible models in the SAT example (with Alaska excluded). Such a display would not be used in practice, but it illustrates how sequential techniques find a good-fitting model even though they search through only a handful of all possible models.

The models on the display are labeled with one-character symbols associated with each explanatory variable. The model labeled 1 has only the intercept. The model labeled T has the intercept and t (the log of the percentage of takers) only. The model labeled ET has the intercept, the log of percentage of takers (t), and state expenditure (e). And so on. All models in the same row contain the same number of explanatory variables. Models are coded so that the variables whose F -statistics are greater than 4 are listed in uppercase letters, and those with F -statistics less than or equal to 4 are listed in lowercase letters. So, for example, in the fit to the four-variable model with income, years, public, and expenditure, listed as $IYPe$, expenditure is not significant, but the other variables are. The lines connecting the boxes in the display represent the paths for getting from one model to another that has one fewer or one more explanatory variable.

Forward Selection

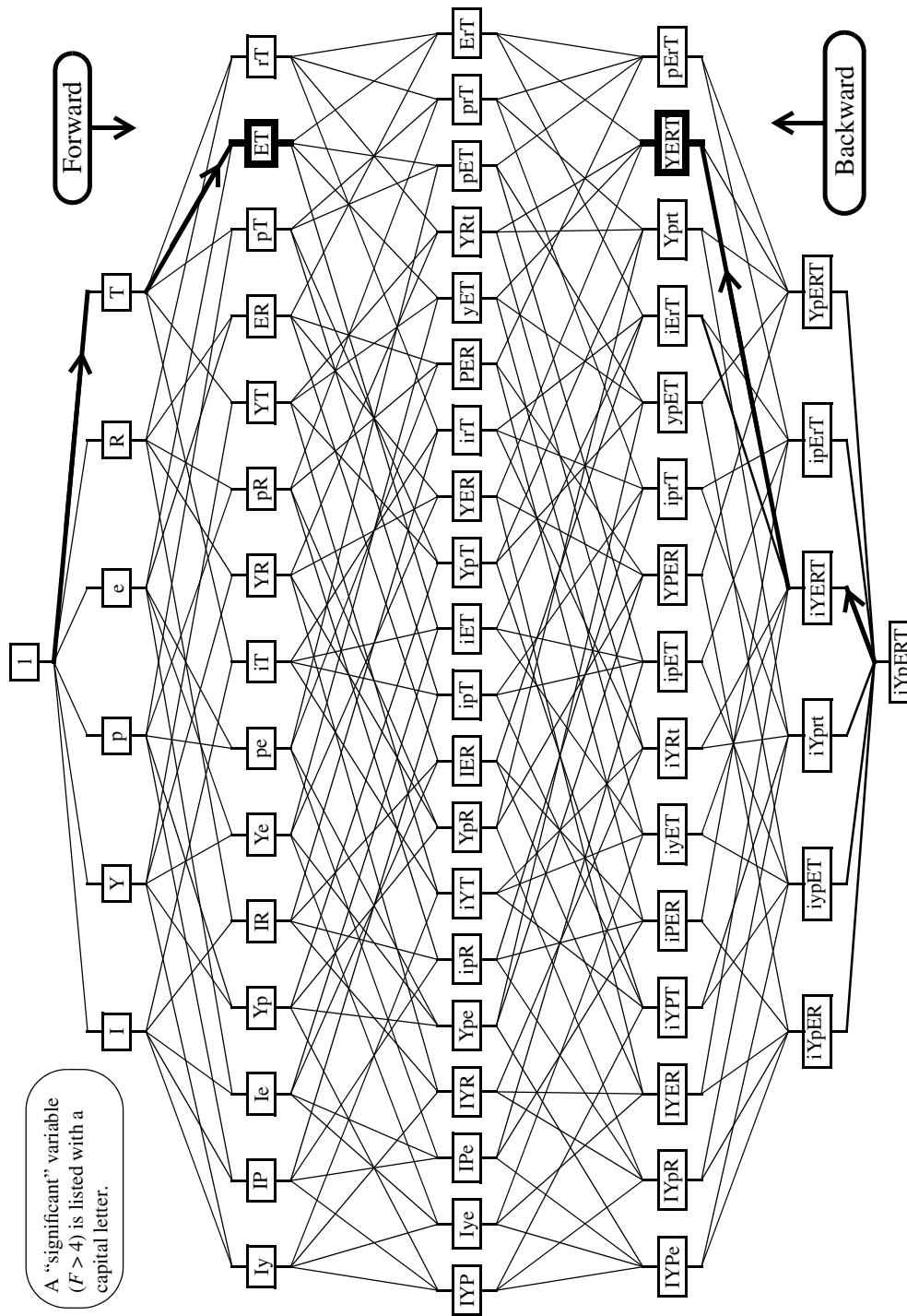
Model 1 is taken as the initial current model, and all single-variable models are considered. In the second row of Display 12.6, models I , Y , R , and T all have coefficients whose F -statistics are larger than 4. The one of these with the largest F -to-enter is T , so it is taken as the next current model. Next, all models that include T and one other variable are examined. Only Y and E add significantly when added to the model with T . Of these, E has the larger F -to-enter, so the next current model is ET . Next, all three-variable models that include E and T are considered, but in no case is the additional variable significant. Therefore, the forward selection process stops, and the final model is ET . The entire procedure required the fitting of only 16 of the 64 possible models.

Backward Elimination

The model with all variables, $iYpERT$, is taken as the starting point. Of the variables in the model, p (public) has the smallest F -statistic, and p is dropped because its F is less than 4. The resulting model is $iYERT$. The least significant coefficient in the new current model is the one for i (income), and its F -statistic is less than 4, so it is dropped. The resulting model is $YERT$. All of the variables in this model have F -statistics greater than 4, so this is the final model. This procedure required the fitting of only 3 of the 64 possible models, and none of the 3 was in the set considered by forward selection.

DISPLAY 12.6

Anatomy of sequential variable selection—SAT example



Stepwise Regression

Stepwise regression starts with one step of forward selection to arrive at the model T . No variables can be dropped, so it tries to add another. It arrives at ET . It then checks whether any of the variables in this model can be dropped. None can, so it seeks to add another variable, but no others add significantly, so the procedure stops at ET .

12.3.5 Compounded Uncertainty in Stepwise Procedures

The cutoff value of 4 for the F -statistic (or 2 for the magnitude of the t -statistic) corresponds roughly to a two-sided p -value of less than 0.05. The notion of “significance” cannot be taken seriously, however, because sequential variable selection is a form of data snooping.

At step 1 of a forward selection, the cutoff of $F = 4$ corresponds to a hypothesis test for a single coefficient. But the actual statistic considered is the largest of several F -statistics, whose sampling distribution under the null hypothesis differs sharply from an F -distribution.

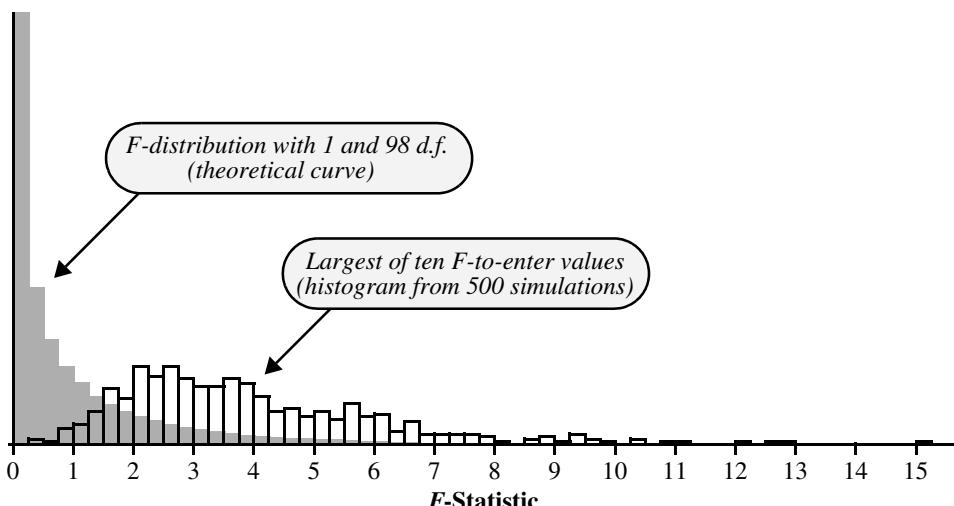
To demonstrate this, suppose that a model contained ten explanatory variables and a single response, with a sample size of $n = 100$. The F -statistic for a single variable at step 1 would be compared to an F -distribution with 1 and 98 degrees of freedom, where only 4.8% of the F -ratios exceed 4. But suppose further that all 11 variables were generated completely at random (and independently of each other), from a standard normal distribution. What should be expected of the largest F -to-enter?

This random generation process was simulated 500 times on a computer. Display 12.7 shows a histogram of the largest among ten F -to-enter values, along with the theoretical F -distribution. The two distributions are very different. At least one F -to-enter was larger than 4 in 38% of the simulated trials, even though none of the explanatory variables was associated with the response.

This and related arguments imply that sequential procedures tend to select models that have too many variables if the set contains unimportant ones. This may not be a serious problem for some data problems, but better techniques are available. In part, variable-selection routines are important for historical reasons, since they were the only tools available when computing time was substantially slower. They are also important because they extend in an obvious way to other types of modeling. Now that computers can look at every possible regression model, however, criteria not based on sequential significance address more pertinent features in desired models.

12.4 MODEL SELECTION AMONG ALL SUBSETS

A second approach to model selection involves fitting all possible subset models and identifying the ones that best satisfy some model-fitting criterion. Since sequential selection is avoided, the criteria can be based directly on the key issues: including

DISPLAY 12.7Simulated distribution of the largest of ten F -statistics

enough explanatory variables to model the response accurately without the loss of precision that occurs when essentially redundant and unnecessary terms are included.

The two criteria presented in this section are the Cp statistic and Schwarz's Bayesian Information Criterion. These are calculated for the fit to a model, so each of the possible regression models receives a numerical score. Before these are examined, some simpler choices should be mentioned.

R^2 and Adjusted R^2

Comparing two models with the same number of parameters is relatively easy: The one with the smaller residual mean square $\hat{\sigma}^2$ is preferred. When applied to comparisons of models that have different numbers of parameters, R^2 (not adjusted) always leads to selecting the model with all variables. Because R^2 is incapable of making a sensible selection, some researchers use the adjusted R^2 . This is equivalent to selecting the model with the smallest $\hat{\sigma}^2$ —a criterion that generally favors models with too many variables.

12.4.1 The Cp Statistic and Cp Plot

The Cp statistic is a criterion that focuses directly on the trade-off between bias due to excluding important explanatory variables and extra variance due to including too many. The bias in the i th fitted value is

$$\text{Bias}\{\hat{Y}_i\} = \mu\{\hat{Y}_i\} - \mu\{Y_i\}.$$

This is the amount by which the mean in the sampling distribution of the i th fitted value differs from the mean it is attempting to estimate. The mean squared error

is the squared bias plus the variance:

$$\text{MSE}\{\hat{Y}_i\} = [\text{Bias}\{\hat{Y}_i\}]^2 + \text{Var}\{\hat{Y}_i\}.$$

The *total mean squared error (TMSE)* for a given model is the sum of these over all observations:

$$\text{TMSE} = \sum_{i=1}^n \text{MSE}\{\hat{Y}_i\}.$$

It is desired to find a subset model with a small value of TMSE. The squared bias terms are small if no important explanatory variables are left out, and the variance terms are small if no unnecessary explanatory variables are included.

The Cp statistic is an estimate of TMSE/σ^2 , based on assuming that the model with all available explanatory variables has no bias. For a model that has p regression coefficients (including β_0), the Cp statistic is computed as

$$C_p = p + (n - p) \frac{(\hat{\sigma}^2 - \hat{\sigma}_{\text{full}}^2)}{\hat{\sigma}_{\text{full}}^2},$$

where $\hat{\sigma}^2$ is the estimate of σ^2 from the tentative model, and $\hat{\sigma}_{\text{full}}^2$ is the estimate of σ^2 from the fit with all possible explanatory variables. One Cp statistic can be calculated for each possible model.

Models with small Cp statistics are looked on more favorably. If a model lacks important explanatory variables, it will show greater residual variability than the full model with all explanatory variables will show. Thus, $\hat{\sigma}^2 - \hat{\sigma}_{\text{full}}^2$ will be large. On the other hand, if the difference in estimates of σ^2 is close to zero, including p in the formula will add a penalty in the Cp statistic for having more explanatory variables than necessary.

The Cp Plot

A Cp plot is a scatterplot with one point for each subset model. The y -coordinate is the Cp statistic, and the x -coordinate is p —the number of coefficients in the model. Each point in the plot is labeled to show the model to which it corresponds, often by including one-letter codes corresponding to each explanatory variable that appears in the model.

The Cp plot, which is available in many statistical computer packages, can be scanned to identify several models with low Cp statistics. The line at $Cp = p$ is often included on the plot, since a model without bias should have Cp approximately equal to p . Some users treat this line as a reference guide for visual assessment of the breakdown of the Cp statistic into bias and variance components; but the Cp statistic itself is the criterion. Thus, the models with smallest Cp statistics are considered. There is nothing magical about the model with the smallest Cp, since the actual ordering is likely to be affected by sampling variability; thus, a different set of data may result in a different ordering. Picking a single model from among all

DISPLAY 12.8

Three criteria for comparing regression models with different subsets of X 's. Subsets that produce small values of these are thought to strike a good balance between small sum of squared residuals (SSRes) and not too many regression coefficients (p).

Criterion to make small	=	SS Res. part	+	Penalty for number of terms
Cp	=	$\frac{SSRes}{\hat{\sigma}_{full}^2} - n$	+	$2p$
BIC	=	$n \times \log\left(\frac{SSRes}{n}\right)$	+	$\log(n) \times (p + 1)$
AIC	=	$n \times \log\left(\frac{SSRes}{n}\right)$	+	$2 \times (p + 1)$

those with small Cp statistics is usually a matter of selecting the most convenient one whose coefficients all differ significantly from zero.

12.4.2 Akaike and Bayesian Information Criteria

The Cp statistic takes a measure of the lack of fit of a model and adds a penalty for the number of terms in the model. As more terms are included in the model, lack of fit decreases, but the penalties increase. Models with small Cp statistics balance the lack of fit and the penalty.

Two alternative model selection statistics incorporating similar penalties are Akaike's Information Criterion (AIC) and Schwarz's Bayesian Information Criterion (BIC). Display 12.8 shows the formulas for these statistics.

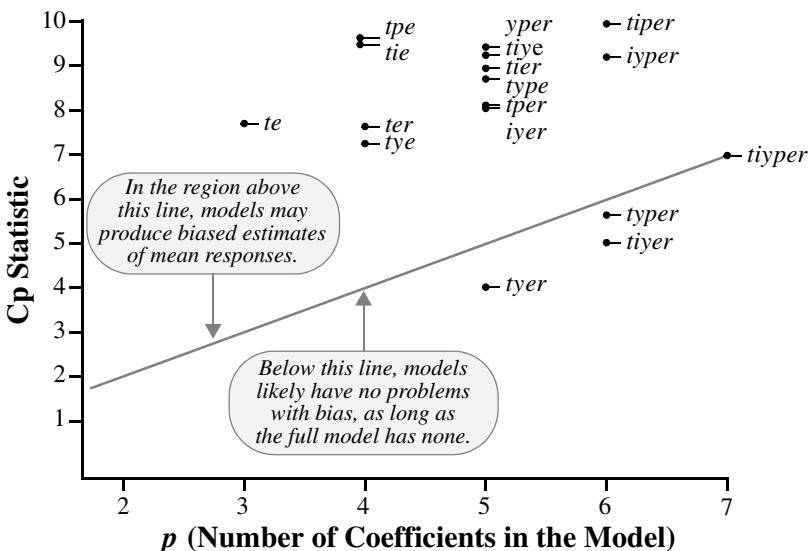
Some computational versions of BIC and AIC differ from the formulas in Display 12.8 by a constant. Since the constant is the same for all models, this will not affect the ordering of models indicated by the model-fitting criteria.

The BIC penalty for number of betas is larger than the AIC penalty for all sample sizes larger than 7, and the difference increases with increasing sample size. Consequently, BIC will tend to favor models with fewer X 's than AIC, especially from large samples. There is no theory for clarifying the relative benefits of these two criteria; they simply reflect different approaches for balancing under- and overfitting. In general, AIC seems more appropriate if there are not too many redundant and unnecessary X 's in the starting set; BIC seems more appropriate if there are many redundant and unnecessary. The distinction between these two situations is rarely clear in practice, but BIC is particularly useful when the researcher places a larger burden of model determination on the computer, such as when a set of measured X 's, their quadratic terms, and all possible two-term interactions are included in the starting set.

In any case, it is important to realize that no criterion can be superior in all situations. It is also important to realize that variable-selection routines will not discover a scientific law of nature. There are usually many possible subsets—formed from highly interconnected explanatory variables—that explain the response

DISPLAY 12.9

Cp plot for state SAT averages (excluding Alaska and showing only those models with $C_p < 10$); $t = \log$ percentage of takers, $i = \text{income}$, $y = \text{years}$, $p = \text{public}$, $e = \text{expend}$, and $r = \text{rank}$



variation equally well. As the famous English statistician George E. P. Box said, “All models are wrong, but some are useful.”

SAT Example

Display 12.9 shows a Cp plot for the SAT data (without Alaska), including only models with C_p less than 10. The models are identified on the plot by the variables included, with $t = \log$ percentage of takers, $i = \text{income}$, and so on. Observe that the Cp statistic for the general model that includes all variables—*tiyper*—equals the maximum number of terms, 7 (including the constant). The model *tyer*, which was identified by backward elimination, has the smallest Cp statistic.

The model *tyer* has the smallest BIC, at 322.4. It is followed by *te* with BIC = 322.7, *tye* with BIC = 324.2, and *ter* with BIC = 324.6. The models with smallest AIC values are *tyer* with AIC = 311.1, *tiyer* with AIC = 311.9, *typer* with AIC = 312.7, and *tiyper* with AIC = 313.9.

12.5 POSTERIOR BELIEFS ABOUT DIFFERENT MODELS

One branch of statistical reasoning, called *Bayesian statistics*, assumes that beliefs in various hypotheses can be expressed on a probability scale. The elegant theorem of English mathematician Thomas Bayes (1702–1761), known as Bayes’ Theorem, specifies precisely how beliefs held prior to seeing the data should be altered by

statistical evidence to arrive at a set of posterior beliefs. In the present setting, Bayesian statistics can be used to update the probability that each subset model is correct, given the available data.

Suppose that consideration is confined to models M_1, M_2, \dots, M_K , and suppose that prior probabilities (probabilities of belief prior to seeing the data) on these models are $\text{pr}\{M_1\}, \text{pr}\{M_2\}, \dots, \text{pr}\{M_K\}$. Applying a Bayesian formulation, Schwarz showed that the updated probabilities for the models, given the data (D), were approximately

$$\text{pr}\{M_j|D\} = \text{pr}\{M_j\} \times \exp\{-\text{BIC}_j\}/\text{SUM}$$

where

$$\text{SUM} = \sum_{i=1}^K \text{pr}\{M_i\} \times \exp\{-\text{BIC}_i\},$$

where $\text{pr}\{M_j|D\}$, called a *posterior probability*, is the probability of M_j after seeing the data. When no prior reasons exist for believing that one model is better than any other, $\text{pr}\{M_j\}$ is the same for all j , and the posterior probabilities are simple functions of the BIC values for each model.

For the SAT example, the values of BIC have been calculated for each model and then used in this formula to identify the posterior probabilities. The posterior probability that the model is *te* is 0.764, followed by *tyer* with 0.115, *tie* with 0.060, and *tpe* with 0.041. The posterior probability that the model is any of the other 60 models is 0.020, and no other single model has a posterior probability greater than 0.006.

12.6 ANALYSIS OF THE SEX DISCRIMINATION DATA

To investigate male/female differences in beginning salaries requires a strategy that convincingly accounts for variables other than the sex indicator before adding it to the equation.

Initial investigation indicates that beginning salary should be examined on the log scale. Scatterplots of log beginning salary versus other variables show that the effect of experience is not linear. Beginning salaries increase with increasing experience up to a certain point; then they level off and even drop down for individuals with more experience. A quadratic term could reproduce this behavior well. A similar effect is seen with age: a correlation between age and experience may be responsible for that effect, but it is wise initially to consider a squared term for both variables. It may seem that seniority (number of months working with the company) is an inappropriate term to include for modeling beginning salary, but its inclusion accounts for increasing beginning salaries over time. Rather than as seniority, it might be thought of as time prior to March 1977 that the individual was hired.

Such considerations point to the necessity of including in the investigation a model rich enough to be extremely unlikely to miss an important relationship between log beginning salary and all explanatory variables other than the sex indicator. A saturated second-order model that includes all quadratic terms and all interaction terms satisfies that criterion. (See Section 12.7.3 for further discussion of second-order models.)

Notation

Fourteen variables appear in the saturated second-order model, so naming variables with single letters utilizes space well, even though it sacrifices clarity. The full complement of explanatory variables is given in Display 12.10. A word represents a model whose explanatory variables consist of the letters that make up the word. For example, the word *saebc* represents the model with seniority, age, education, the square of age, and the product of age with education—a model with six parameters in all (including the constant).

DISPLAY 12.10

Explanatory variables for the sex discrimination data

Main effect variables	Quadratic variables	Interaction variables
$s = \text{Seniority}$	$t = s^2$	$m = s \times a$
$a = \text{Age}$	$b = a^2$	$c = a \times e$
$e = \text{Education}$	$f = e^2$	$n = s \times e$
$x = \text{Experience}$	$y = x^2$	$k = a \times x$
		$v = s \times x$
		$q = e \times x$

A total of $16,384 (= 2^{14})$ models are formed by all possible combinations of the 14 terms in the preceding box. The goal is to find a good subset of these terms for explaining log salary, to which the sex indicator variable may be added. Although quadratic and interaction terms should be considered, it is not good practice to include quadratic terms without the corresponding linear term; neither is it good practice to include interaction terms without the two corresponding main effects. The preferred approach, therefore, is to identify the better models (of the 16,384) and then to identify the best subset of these that meets the narrower criteria restrictions of “good practice.”

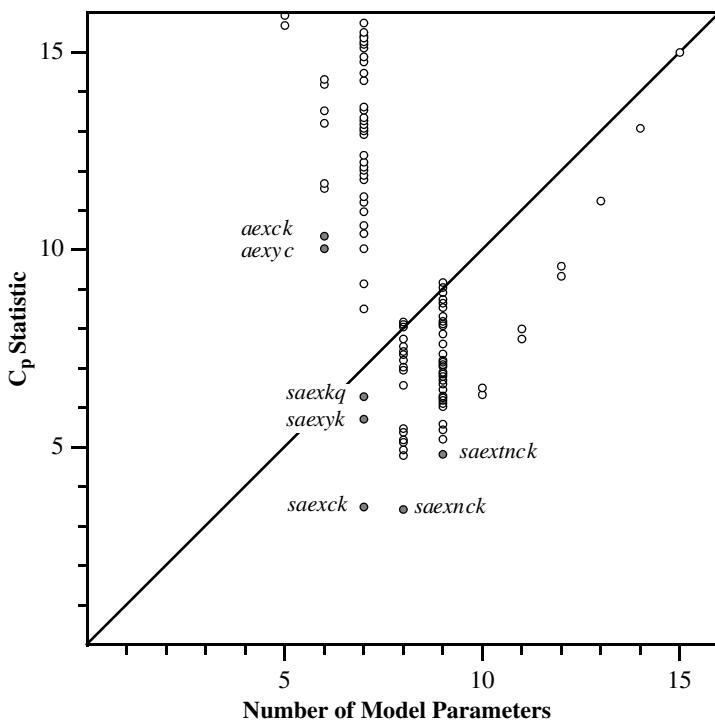
Identifying Good Subset Models

Display 12.11 shows a Cp plot. The Cp statistics were based on the full model containing all 14 terms. To reduce clutter, the plot displays Cp statistics only for models that meet the criteria discussed in the previous paragraph and, in addition, have small Cp values. Only a few good-fitting models with more than seven parameters are shown on this plot.

Models *saexck* and *saenck* have lower Cp statistics than any others, and they appear to have no bias relative to the saturated model (since they fall near or below the line at $Cp = p$). The BIC clarifies the issue further: among 382 models examined, the BIC assigns a posterior probability of 0.7709 to *saexck*, with *saexyc*

DISPLAY 12.11

Cp plot for the sex discrimination study



receiving the second highest posterior probability of 0.0625. Display 12.12 shows the 20 models that recorded the highest posterior probabilities.

Evaluating the Sex Effect

When the sex indicator variable is added into model *saexck*, its estimated coefficient is -0.1196 , with a standard error of 0.0229 . The *t*-statistic is 5.22 , and the test that the coefficient is zero against the alternative that it is negative has a one-sided *p*-value $= 6.3 \times 10^{-7}$. Thus, conclusive evidence exists that sex is associated with beginning salary, even after seniority, age, education, and experience are accounted for in this model. The median female salary is estimated to be $\exp(-0.1196) = 0.887$ times as large as (that is, about 11% less than) the median male salary, adjusted for the other variables.

Accounting for Model Selection Uncertainty Using the BIC

The preceding inferential statements pertain specifically to the single model *saexck*, selected to represent the influence of the potentially confounding variables. One criticism of this result is that it fails to account for the uncertainty involved in the model selection process. A second criticism is that the addition of the sex indicator

DISPLAY 12.12 Bayesian posterior analysis of the difference between male and female log beginning salaries

Model	BIC	Posterior probability	Coefficient of sex	SE	One-sided p-value
saexck	-404.15	0.6909	-0.1196	0.0229	6.16E-07
saexnck	-401.91	0.0738	-0.1173	0.0229	9.31E-07
saexyc	-401.63	0.0558	-0.1287	0.0226	8.21E-08
saexkq	-401.03	0.0306	-0.1244	0.0221	1.16E-07
saexbck	-400.32	0.0150	-0.1195	0.0229	6.57E-07
saextck	-400.21	0.0134	-0.1189	0.0232	8.97E-07
saexbkq	-400.19	0.0132	-0.1206	0.0221	2.36E-07
saexync	-400.16	0.0129	-0.1258	0.0225	1.34E-07
saexfek	-399.95	0.0104	-0.1196	0.0230	6.80E-07
saexckq	-399.89	0.0097	-0.1208	0.0230	5.43E-07
saexyck	-399.74	0.0084	-0.1257	0.0232	2.75E-07
saexmck	-399.68	0.0079	-0.1195	0.0231	7.32E-07
saexvck	-399.63	0.0076	-0.1196	0.0231	7.17E-07
aexyc	-399.49	0.0065	-0.1247	0.0238	5.48E-07
aexck	-399.18	0.0048	-0.1135	0.0246	6.86E-06
saexyq	-398.64	0.0028	-0.1328	0.0218	1.46E-08
saextyc	-398.42	0.0023	-0.1271	0.0228	1.41E-07
saectnck	-398.14	0.0017	-0.1163	0.0232	1.41E-06
saexbc	-397.95	0.0014	-0.1230	0.0237	6.83E-07
saexyvc	-397.93	0.0014	-0.1281	0.0226	9.97E-08
all others combined		0.0294			

$s = \text{Seniority}$	$t = s^2$	$m = s \times a$	$c = a \times e$
$a = \text{Age}$	$b = a^2$	$n = s \times e$	$k = a \times x$
$e = \text{Education}$	$f = e^2$	$v = s \times x$	$q = e \times x$
$x = \text{Experience}$	$y = x^2$		

variable to other good-fitting models might produce less convincing evidence of a sex effect.

The posterior probabilities for various models allow the analyst to address both criticisms. Suppose that δ represents the coefficient of the sex indicator variable and that $f(\delta|M_j)$ represents some function $f(\delta)$ based on the assumption that model M_j is correct. The Bayesian analysis permits an average estimate of $f(\delta)$ over different models. Furthermore, the average is weighted to represent the strength of belief in each of the various models. The estimate of $f(\delta)$ in this way incorporates the uncertainty in model selection. The weighted average is

$$f\{\delta|D\} = \sum_{i=1}^K f(\delta|M_i) \times \text{pr}\{M_i|D\}.$$

Two choices for $f(\delta)$ are δ itself (the estimate of the sex effect) and the p -value for testing its significance. The sex indicator variable was added to each of the 382 models and combined as shown in the preceding equation to produce the single (posterior) estimate of -0.1206 and a single (posterior) p -value of 6.7×10^{-7} (this requires some programming). A listing of the top 20 candidate models, together with their contributions to the posterior information, is shown in Display 12.12.

The use of Bayesian statistics requires that somewhat different interpretations be given to the estimate and to the p -value. The average estimate is termed a *Bayesian posterior-mean estimate*, while the average p -value is called the *posterior probability* that δ is greater than zero. The probability refers to a measure of belief about the hypothesis that the sex coefficient is zero. This is a different philosophical attitude than any previously considered, but two issues argue in its favor for this application. First the formula itself gives the p -value for every model a chance to contribute in the final estimate, with a weight appropriate to the support given the model by the data. Second, no known statistical procedures based on other philosophies incorporate model selection uncertainty in this way.

12.7 RELATED ISSUES

12.7.1 The Trouble with Interpreting Significance when Explanatory Variables Are Correlated

When explanatory variables are correlated, interpretation of individual regression coefficients and their significance is quite difficult and perhaps too convoluted to be of use. The central difficulty is that interpreting a single coefficient as the amount by which the mean response changes as the single variable changes by one unit, holding all other variables fixed, does not apply; this is because the data generally lack experience with situations where one variable can be changed in isolation.

Consider the variable *expend*—the per student state expenditure—in the SAT example. This variable can be added to any model that does not already contain it. This was done for all models not containing *expend*, with the results shown in Display 12.13.

The first three columns of Display 12.13 are self-explanatory, the key point being that the p -values vary widely. The last column is the addition to R^2 resulting from the addition of *expend* to the model. The variation in these percentages can be traced to the correlation between *expend* and the other variables used in the model. The last column can be interpreted as the extra sum of squares due to *expend*, expressed as a percentage of the total sum of squares. The fourth column is the extra sum of squares due to *expend*, expressed as a percentage of the sum of squared residuals. The huge variations in that column arise because of the differing amounts of variability explained by the different models not containing *expend*. This is most directly related to the variation in p -values for *expend*, but the correlational aspect also plays a role.

Ultimately, the answer to the often-asked question, “Which variables are significant?” has to be “It depends.” Similarly, there is usually no definitive answer

DISPLAY 12.13 Statistical measures for the contribution of *expend* to different models (Alaska removed)

Model	<i>t</i> -statistic	Two-sided <i>p</i> -value	% Variation explained by <i>expend</i>	
			Previous residual	Total variation
<i>E</i>	0.27	0.79	0.2	0.2
<i>I+E</i>	0.40	0.69	0.3	0.2
<i>Y+E</i>	0.86	0.39	1.6	1.4
<i>P+E</i>	0.18	0.86	0.1	0.1
<i>R+E</i>	4.78	0.00002	33.2	7.4
<i>T+E</i>	6.04	0.0000002	44.3	8.4
<i>IY+E</i>	0.07	0.94	0.0	0.0
<i>IP+E</i>	0.11	0.92	0.0	0.0
<i>IR+E</i>	4.88	0.00001	24.6	6.7
<i>IT+E</i>	5.88	0.0000005	43.4	8.1
<i>YP+E</i>	1.05	0.30	2.4	2.1
<i>YR+E</i>	4.20	0.0001	28.2	4.3
<i>YT+E</i>	5.31	0.000003	38.5	6.3
<i>PR+E</i>	6.02	0.0000003	44.6	9.4
<i>PT+E</i>	5.91	0.0000004	43.7	8.2
<i>RT+E</i>	6.13	0.0000002	45.5	8.4
<i>IYP+E</i>	0.92	0.36	1.9	0.9
<i>IYR+E</i>	4.35	0.00008	30.1	4.2
<i>IYT+E</i>	5.20	0.000005	38.0	6.2
<i>IPR+E</i>	5.57	0.000001	41.4	8.0
<i>IPT+E</i>	5.59	0.000001	41.5	7.5
<i>IRT+E</i>	5.92	0.0000004	44.4	7.9
<i>YPR+E</i>	4.80	0.00002	34.4	5.3
<i>YPT+E</i>	4.78	0.00002	34.2	5.1
<i>YRT+E</i>	5.23	0.000005	38.3	5.6
<i>PRT+E</i>	6.27	0.0000001	47.2	8.8
<i>IYPR+E</i>	4.27	0.0001	29.8	4.1
<i>IYPT+E</i>	4.33	0.00009	30.4	4.3
<i>IYRT+E</i>	5.00	0.00001	36.8	5.1
<i>IPRT+E</i>	5.94	0.0000004	45.1	8.0
<i>YPRT+E</i>	5.13	0.000007	38.0	5.4
<i>IYPRT+E</i>	4.64	0.00003	33.9	4.5

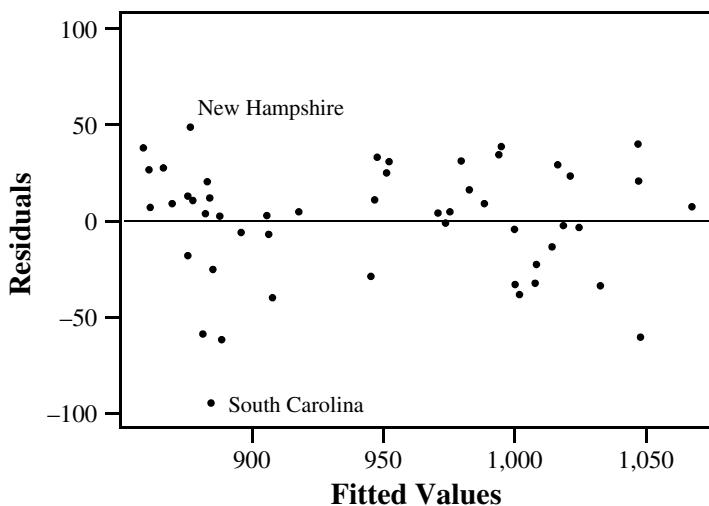
to the question, “Which variable is most significant?” It is best to try to focus attention on questions that ask specifically for the assessment of significance after accounting for other specified variables or groups of variables.

12.7.2 Regression for Adjustment and Ranking

One valuable use of regression is for adjustment. Consider the regression of state SAT averages on percentage of eligible students in the state who decide to take the test and on the median class rank of the students who take the exam. A residual plot from the fit (excluding Alaska) is shown in Display 12.14. The residuals are the

DISPLAY 12.14

Scatterplot of residuals versus fitted values from the regression of state SAT average on percentage of takers and median class rank of takers



state SATs with the estimated linear effects of log percentage of takers and median class rank removed. Thus, they serve as the SATs adjusted for these two variables (and are scaled so that the average is zero). New Hampshire, for example, whose “raw” SAT average is 925 (which is 23 points less than the overall average of all states) has a very high percentage of eligible students who take the exam and a fairly low median class rank among the takers. The estimated mean SAT score for a state with these values is 876. Since New Hampshire’s actual SAT score was 925, its value exceeded the prediction (according to the model with percentage of takers and median class rank) by 49 points. It is therefore appropriate to say that New Hampshire’s SAT adjusted for percentage of takers and median class rank of takers is 49 points above average.

The ranking on the left side of Display 12.2 is based directly on the residuals shown in Display 12.14. New Hampshire has the largest residual, and therefore the largest SAT adjusted for percentage of takers and median class rank. South Carolina has the smallest.

The ranking on the right side of Display 12.2 is based on the residuals from a model that also includes expenditure. In this case, New Hampshire’s SAT, adjusted for percentage of takers, median class rank, and expenditure, is 67 points above the average predicted for a state with the same values of percentage of takers, median rank, and expenditure.

12.7.3 Saturated Second-Order Models

A saturated second-order model (SSOM) includes the squares of all explanatory variables and all cross products of pairs of explanatory variables. It describes $\mu(Y)$

as a completely arbitrary parabolic surface. The SSOM should contain a subset model that describes the regression surface well.

Numbers of Subset Models

Given K original variables, the SSOM itself contains $[K(K + 3)/2] + 1$ parameters. When all possibilities are taken into account, the total number of p -parameter hierarchical models (meaning ones where, for example, X_1^2 appears only if X_1 also appears, and X_1X_2 appears only if both X_1 and X_2 appear) available from K original variables is

$$\sum_{j=0}^{K} C_{K,j} \times C_{(C_{j+1,2}), (p-1-j)},$$

where $C_{n,m}$, recall, is $n!/[m!(n - m)!]$ and $C_{n,m} = 0$ whenever $n < m$. Holding out the sex indicator in the sex discrimination problem leaves $K = 4$ explanatory variables: *seniority*, *age*, *education*, and *experience*. The full SSOM has 15 parameters, and subset models have 1, 2, 3, ..., 14 parameters. The number of distinct hierarchical models that can be considered is the number of hierarchical modes with one parameter plus the number with two parameters and so on up to the number of hierarchical models with fourteen parameters. According to the formula, a total of 1,337 distinct models can be considered.

Strategies for Exploring Subsets of the SSOM

The SAT study featured $K = 6$ explanatory variables. The SSOM for that problem contains 28 parameters. The number of hierarchical models with 17 parameters, for example, is 352,716, and the total of all hierarchical models to consider is 2,104,489. Things rapidly get out of hand, and it becomes necessary to plan a strategy for sorting through some—but not all—of the models for promising candidates.

The strategy employed in the sex discrimination study was to examine all subset models up to some level ($p = 7$) first. Forward selection from the best models with $p \leq 7$ came next, followed by backward elimination from the saturated second-order model. Finally, models in the neighborhood of promising models were included. A *neighborhood* of one model consists of all models that can be obtained by adding or dropping one variable.

Another strategy involves grouping variables into sets that have a common theme and exploring each set separately to determine its best subset. When these are combined into an overall model, subsets consisting of products of retained variables from one set with those from another may also be examined.

A third strategy is to identify which of the original variables are important, by examining all subsets with main effects only (i.e., excluding squared and interaction terms), and then to build the SSOM from those variables alone and explore subsets.

12.7.4 Cross Validation

Inference after the use of a variable selection technique is tainted by the data snooping involved in the process. The selected model is likely to fit much better to

the data that gave it birth than to fresh data. Thus, *p*-values, confidence intervals, and prediction intervals should be used cautiously.

If the data set is very large, the analyst may benefit from dividing it at random into separate model construction and validation sets. A variable selection technique can be used on the model construction set to determine a set of explanatory variables. The selected model can then be refit on the validation set, without any further exploration into suitable explanatory variables, and inferential questions can be investigated on this fit, ignoring the construction data. When the purpose of the regression analysis is prediction, it is recommended that the validation data set be about 25% of the entire set. Although the saving of the 25% of the data for validation seems reasonable for other purposes, the actual benefits are not very well understood.

12.8 SUMMARY

Model selection requires an overall strategy for analyzing the data with regression tools (see Display 9.9). After giving initial thought to a game plan for investigating the questions of interest, the preliminary analysis consists of a combination of exploration, model fitting, and model checking. Once some useful models have been identified, the answers to the questions of interest can be addressed through inferences about their parameters.

Tools for initial exploration include graphical methods (such as a matrix of scatterplots, coded scatterplots, jittered scatterplots, and interactive labeling) and correlation coefficients between various variables. Certain tricks for modeling are used extensively in the case studies—indicator variables, quadratic terms, and interaction terms. For model checking and model building, a number of other tools are suggested: residual plots, partial residual plots, informal tests of coefficients, case influence statistics, the Cp plot, the BIC, and sequential variable selection techniques. Finally, some inferential tools are presented: *t*-tests and confidence intervals for individual coefficients and linear combinations of coefficients, extra sums-of-squares *F*-tests, prediction intervals, and calibration intervals.

SAT Study

Although this example is used to demonstrate variable-selection techniques, the actual analysis is guided by the objectives, and variable selection played only a minor role. These data have been used to rank the states on their success in secondary education, but in this regard selection bias poses a serious problem. In some states, for example, a high SAT average reflects the fact that only a small proportion of students—the very best ones—took the test, rendering the self-selected sample far from representative of high school students in the state overall. One goal of the regression analysis is to establish a ranking that accounts for this selection bias. Although overcoming the limitations of a self-selected sample is impossible, it is possible to rank the states after subtracting out the effects of the different proportions of students taking the test and their different median class rankings. Accomplishing